Université de Montréal

Méthodologie pour l'analyse de données de criblage: application à l'étude de la leucémie myéloïde aiguë

par Caroline Labelle

Département de biochimie et médecine moléculaire
Faculté de médecine

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en bioinformatique

Avril, 2018

À Anna, Pierre et Antoinette

Résumé

Les expériences de criblage à haut débit (CHD) permettent d'évaluer l'efficacité de composés chimiques pour diverses conditions. L'analyse des données se fait en générant une courbe de type dose-réponse et en interprétant les paramètres de celle-ci. Les outils présentement utilisés pour de telles analyses manquent de flexibilité dans leur protocole et ne proposent aucune approche pour la comparaison d'ajustements. Le but de ce travail est d'établir un processus automatisé capable de générer, analyser et comparer des courbes dose-réponse-réponse, et ce, pour différents protocoles expérimentaux. Les courbes sont obtenues en ajustant le modèle log-logistique à l'aide d'un algorithme de régression non-linéaire. Nous mettons de l'avant le concept de groupe qui permet de faire un ajustement sur les données d'un ensemble d'échantillons. Cette approche semble robuste, même en présence de données aberrantes. Nous proposons aussi une approche statistique utilisant la simulation de données et le ré-échantillonnage pour obtenir des intervalles de confiance et faire la comparaison de deux ajustements. Le processus proposé est appliqué dans un contexte expérimental qui étudie les effets de composés sur des cellules leucémiques. Bien que plusieurs analyses se font en interprétant principalement l'IC_{50}, les résultats obtenus suggèrent que d'autres métriques devraient aussi être analysée pour déterminer l'efficacité d'un composé. Ces différentes métriques aident à bien caractériser et mieux comprendre un composé chimique.

Mots clés : Régression non-linaire, Simulation Monte-Carlo, Ré-échantillonnage Bootstrap, Criblage à haut débit, Leucémie myéloïde aiguë, Bio-informatique

Abstract

High throughput screens (HTS) allow us to evaluate the efficacy of multiple chemical compounds for various conditions. HTS data is analyzed by fitting a mathematical model to normalized dose-response data and by interpreting the adjusted parameters. The tools currently available to do such analysis lack flexibility in their protocol and are not able to statistically compare two fittings. The aim of this project is to establish an automated pipeline able to generate, analyse and compare fittings for various experimental contexts. The fittings are obtained by adjusting the parameters of the log-logistic model with a non-linear regression algorithm. We are putting forward the concept of a group, which is used to fit the model for a given set of samples. This approach seems to be robust even when the data contains outliers. We also propose a statistical approach based on data simulation and resampling to obtain confidence intervals for the adjusted parameters and to compare two independant fittings. Our proposed pipeline is applied to experimental context where we study the effects of chemical compounds on leukemic cells. While many analysis are done by only interpreting the IC_{50}, our results suggest that other parameters should also be considered when establishing th efficacy of a compound. These parameters help to characterize and better understand a given chemical compound.

Keywords: Nonlinear regression, Monte-Carlo simulation, Bootstrap resample, High throughput screen, Acute myeloid leukemia, Bioinformatics

Résumé de Vulgarisation

Le processus de développement de nouveaux médicaments et de nouvelles thérapies comporte plusieurs étapes. Dans un premier temps, l'efficacité de diverses molécules chimiques est évalué dans différentes conditions. L'efficacité d'une molécule est déterminée en analysant ces effets sur des cellules d'intérêt et ce, à différentes concentrations. Les réponses cellulaires obtenues sont alors utilisées pour ajuster un modèle mathématique de telle sorte à générer une courbe illustrant la relation dose-réponse. Une fois le modèle ajusté aux données expérimentales, il est possible d'interpréter ces différents paramètres en tant que descriptifs de l'efficacité de la molécule étudiée. Le présent travaille vise à établir un processus automatisé pour faire de telles analyses. Notre approche est plus flexible que celle proposée par les divers outils couramment utilisés : elle permet de faire l'analyse des effets d'une molécule sur les cellules d'un seul patient, comme sur les cellules d'un ensemble de patients (nous introduisons la notion de groupe). Cette approche semble robuste, même en présence de données aberrantes, et peut être très informative. De plus, nous introduisons une méthodologie de comparaison statistique. Il est alors possible de comparer l'efficacité de composé chimique dans différents contextes, selon différentes métriques. Le processus proposé est appliqué dans un contexte expérimental qui étudie les effets de composés sur des cellules leucémiques. Bien que plusieurs analyses se basent principalement sur la concentration nécessaire pour générer 50% de la réponse maximale, les résultats obtenus suggèrent que d'autres métriques devraient aussi être analysée pour déterminer l'efficacité d'un composé. Ces différentes métriques aident à bien caractériser et mieux comprendre un composé chimique.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

CHD Criblage à haut débit

CLS Cellules leucémiques souches

LMA Leucémie myéloïde aiguë

nM Nanomole

PAD Processus d'ajustement par défaut

RB Ré-échantillonage Bootstrap

RC Rémission complète

SMC Simulation Monte-Carlo

Chapitre 1

Introduction

La découverte de médicaments est un processus hautement multidisciplinaire qui englobe, entre autre, diverses branches de la biologie, les domaines de la chimie, de l'informatique et des mathématiques [1]. Par la passé, la découverte se faisait en identifiant l'élément actif d'une plante ou d'un minéral naturel. Les cibles thérapeutiques, c'est-à-dire les entités moléculaires chez l'humain qui intéragissent avec un composé chimique, n'étaient étudiées qu'une fois l'élément actif confirmé. De nos jours, le processus de découverte de médicament se fait plutôt à l'inverse : une cible pertinente est premièrement identifiée, puis différentes expériences sont mises en place pour analyser l'activité de cette cible dans diverses conditions [2]. Une telle approche permet de déployer des efforts de recherche de façons perninente et précise, ainsi que dans un contexte où il y a demande et nécessité d'une nouvelle approche thérapeutique.

Les avancements technologiques et biomédicales des dernières années ont profité à la découverte de plusieurs nouvelles cibles thérapeutiques et à l'accélération du processus de découverte de médicament [2]. Le criblage à haut débit (CHD) permet de charactériser de façons qualitative et quantitave un très grand nombre de composés chimiques (jusqu'à 10 000 composés par jour) dans un contexte *in vitro* et ce, rapidement [2, 3, 4]. Le CHD permet aussi l'élimination rapide de composés inaptes dans le contexte d'une étude précise

[2, 5]. Différents composés sont alors mis individuellement en contact avec des cellules d'intérêt dans les puits d'une plaque multi-puits. Une seule concentration est testée, celle-ci étant généralement élévée. Les composés générant les réponses cellulaires les plus importantes sont alors libellés comme étant des "*hits*" [4, 6]. Ces *hits* ont alors le potentiel de mener à l'identification et au développement d'un nouveau médicament ou d'une nouvelle thérapie.

L'efficacité de ces *hits* est alors validées par expérience de criblage de type dose-réponse [3]. Chaque composé est mis en contact avec des cellules d'intérêt et ce, pour une série de concentrations. L'ensemble des réponses cellulaires est utilisé pour la modélisation d'une courbe dose-réponse à partir de laquelle il est possible d'extraire des métriques descriptives de l'efficacité du composé [6, 7, 8]. Une de ces métriques est l'IC_{50}. En chimie, cette valeur se définit comme étant la concentration nécessaire pour générer une réponse équivalant à 50% de la réponse asymptotique maximale [9]. Pour les mêmes conditions, nous cherchons à identifier les composés ayant un faible IC_{50} [8], soit les composés capables de générer rapidement une réponses maximale chez les cellules d'intérêt. Les composés sélectionnés lors de ce deuxième filtrage sont libellés comme étant des "*leads*" [3, 4]. Les structures de ces *leads* seront par la suite optimisées par des chimistes dans le but d'obtenir un composé qui pourra éventuellement passer en test clinique [3].

Le présent mémoire porte sur l'analyse de l'efficacité d'un composé, et plus précisement sur l'analyse des données de CHD de type dose-réponse. Cette phase de la découverte de médicament relève principalement des domaines de l'informatique et des mathématiques. Les sections qui suivent présenteront les concepts clés de cette analyse, soit les modèles mathémiques générant une courbe dose-réponse ainsi que les algorithmes optimisant l'ajustement de tels modèles.

1.1 Courbe dose-réponse

Les données obtenues par CHD de type dose-réponse sont utilisées pour ajuster un modèle mathématique duquel des métriques d'intérêt peuvent par la suite être extraites. La présente séction traite des différents modèles qui peuvent être employés dans ce contexte. La section qui suit traitera des algorithmes utilisés pour l'ajustement de ces modèles.

Les modèles de dose-réponse prennent tous la forme générale $y = f(x; \theta)$ où θ représente l'ensemble des paramètres du modèle, x une concentration et y une réponse [10, 11]. "Paramètres" fait ici référence à des constantes inconnus qui seront estimées lors de l'ajustement. Ces estimations seront alors représentatives des données expérimentales de CHD utilisées.

La courbe dose-réponse est caractérisée par deux quasi-plateaux aux asymptotes supérieure et inférieure. Ces asymptotes sont estimées à partir des réponses maximale et minimale obtenues expérimentalement. La courbe se définit aussi par son tracé sigmoïde. Le point d'inflexion déterminant ce changement de courbature peut d'ailleurs souvent être interprété comme étant l'IC_{50}, tel que défini plus haut (Fig. 1.1 p.5) [12].

La forme de la courbe dépend grandement de la distribution des données à partir desquelles l'ajustement est fait. Il est important, lors de la planification du protocol expérimental, de choisir un ensemble de concentrations capable de générer des réponses diversifiées [12]. Idéalement, les concentrations testées s'étaleraient sur au moins deux ordres de magnitude. Cela augmente ainsi les probabilités de couvrir l'ensemble des réponses, des plus minimes aux plus importantes. Le nombre de concentrations est aussi à considérer. Pour un modèle mathématique à quatre paramètres, un minimum de quatre concentrations est nécessaire : il ne devrait jamais y avoir plus de paramètres que de concentrations. Il est recommandé d'utiliser entre cinq et sept concentrations [11]. Cela étant dit, le nombre de concentrations utilisées suit le principe des rendements non-proportionnels [13] : utiliser un très grand nombre de concentrations augmentera significativement les coûts d'expérience

sans pour autant améliorer la qualité de l'ajustement.

De plus, le sens de la courbe est lui aussi dépendant des données utilisées. Une courbe dose-réponse peut être ascendante ou descendante selon le type de réponses étudiées. Les exemples et analyses contenus dans ce travail sont tous faits dans un contexte où la réponse étudiée augmente avec les concentrations. Les différences mathématiques entre ces deux contextes seront brièvement expliquées. De plus, pour faciliter l'analyse et la compréhension des résultats expérimentaux, nous avons choisi, moi et mes collaborateurs, de travailler avec le logarithme en base 10 pour les valeurs des concentrations. Les équations présentées réflètent cette désicion.

Le modèle à ajuster doit être représentatif de la tendance globale du jeu de données et doit donc être sélectionné sur des bases théoriques et empiriques [14]. Il existe différents modèles pour les analyses de type dose-réponse, mais le plus communément utilisé reste le modèle log-logistique [15, 14, 10, 11].

Modèle log-logistique. La modélisation log-logistique fut principalement développée dans le contexte de la rechercher pharmaceutique il y a plus de 60 ans [16]. Bien qu'elle soit premièrement et principalement utilisée pour l'analyse de données de survie [17], la modélisation log-logistique est aujourd'hui utilisée dans de très diverses applications, allant de l'étude des herbicides à l'analyse de l'efficacité d'un composé chimique [17, 15, 11].

Le modèle log-logistique standard comprend quatre paramètres, soit a, b, c et s (Éq. 1.1) (Fig. 1.1, p.5). Ceux-ci représentent les asymptopes inférieure et supérieure, l'abscisse du point d'inflexion et la pente, respectivement.

$$f(x) = a + \frac{b-a}{1+10^{s\cdot(\log_{10} c - \log_{10} x)}} \tag{1.1}$$

Tel que défini par le concept d'asymptote, a et b sont les réponses en l'absence de composé et pour une concentration infinément grande. En l'absence de contrôles positifs

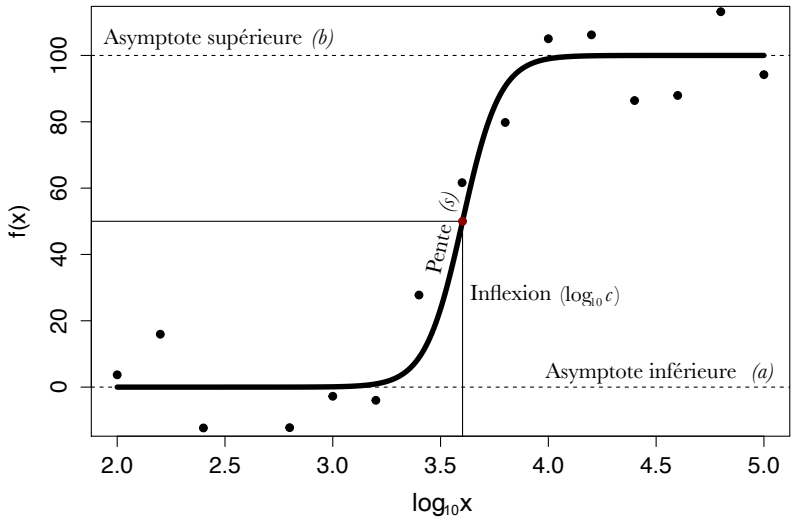


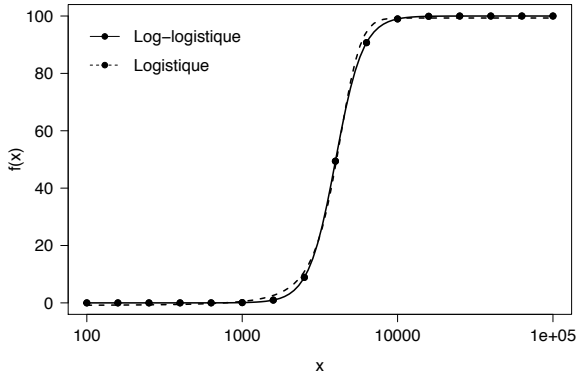
Figure 1.1 – Paramètres du modèle log-logistique. Ajustement du modèle log-logistique à un jeu de données synthétiques tel que représenté par l'ensemble de points noirs. La courbe dose-réponse obtenue depuis l'ajustement est tracée en noir. Les valeurs des asymptotes inférieure et supérieure sont identifiées par les tracés horizontaux hachés. Les coordonnées du point d'inflexion sont quant à elles identifiées par les tracés vertical et horizontal continus.

(réponses pour une très grande concentration) et/ou négatifs (réponses en l'absence de composé), ces valeurs représentent alors des estimations pour les réponses minimales et maximales. Dans le cas où de tels contrôles expérimentaux sont disponibles et utilisés lors de l'ajustement du modèle, les asymptotes sont alors représentatives des réponses minimale et maximale et peuvent être interprétées ainsi. Le paramètre c indique la concentration à laquelle la courbe change de concavité. Le modèle log-logistique étant symétrique autours du point d'inflexion (Fig. 1.3a p.10), c peut aussi être interprété comme étant l'IC_{50}. Finalement, s est la pente instantannée lorsque $x = c$.

L'équation 1.1 décrit une courbe ascendante telle qu'illustrée dans les figures 1.1 et 1.3a. Pour obtenir une courbe descendante, il suffit d'inverser le signe de la pente tel que $s \Rightarrow -s$ ou d'inverser le positionement de c et x tel que $10^{s\cdot(\log_{10} c - \log_{10} x)} \Rightarrow 10^{s\cdot(\log_{10} x - \log_{10} c)}$. Il est intéressant de noter qu'avec une approche algorithmique efficace, l'équation 1.1 sera adéquatement ajustée à des données ayant une tendance descendante : l'estimation de s sera alors négative.

Le modèle log-logistique est souvent référé à tort à l'équation de Boltzman, soit l'équation décrivant le modèle logistique. Les paramètres de ce modèle sont ajustés selon les concentrations réelles plutôt que leur transformation logarithmique. Lorsque les asymptotes d'un jeu de données sont facilement estimables, les deux modèles convergent vers des ajustements semblables (Fig. 1.2a p.7). Cependant, s'il y a absence de plateaux distincts et qu'il ait alors plus difficiles d'estimer les asymptotes, le modèle log-logistique est plus approprié (Fig. 1.2b p.7). Effectivement, dans cette situation le modèle logistique a tendance à converger vers des valeurs qui sont dépourvues de sense, dans le contexte biologique de l'étude (ex : $a = -\infty$) [14]. Une estimation biaisée des limites affectera la valeur du paramètre c (IC_{50}) : l'estimation de ce paramètre s'éloignera de la valeur réelle. Dans le contexte du présent mémoire, nous travaillons au développement d'un processus pour l'analyse de données expérimentales. Celles-ci étant généralement bruitées, il est donc préférable d'utiliser le modèle log-logistique.

(a) Données non-bruitées



(b) Données bruitées

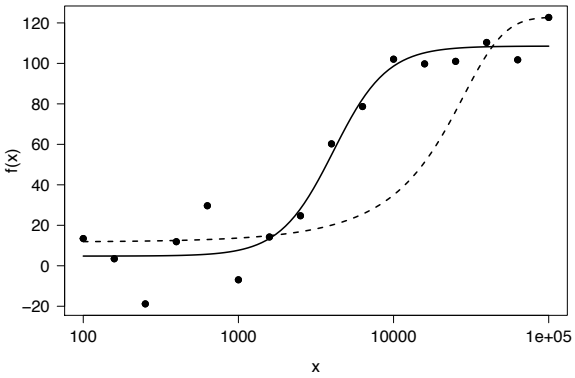


Figure 1.2 – Comparaison des modèles log-logistique et logistique ajustés à différents jeux de données. Les modèles log-logistique et logistique à quatre paramètres sont individuellement ajustés à un jeu de données non-bruitées (1.2a) et à un jeu de données bruitées* ($\sigma = 15$) (1.2b). Les courbes obtenues depuis les ajustements sont représentées par les sigmoïdes noires continues pour le modèle log-logistique, et par les sigmoïdes noires hachées pour le modèle logistique. * *Voir le Chapitre 2 pour la génération des jeux de données synthétiques bruitées.*

Bien que le modèle présenté précédemment ait quatre paramètres, il existe certaines variations. Le modèle log-logistique à trois paramètres (Éq. 1.2) assumme une asymptote inférieure telle que $a = 0.00$.

$$f(x) = \frac{b}{1 + 10^{s \cdot (\log_{10} c - \log_{10} x)}} \tag{1.2}$$

Le modèle log-logistique à deux paramètres (Éq. 1.3) assumme quant à lui que la courbe soit bornée par des aymptotes constantes telles que $a = 0.00$ et $b = 100.00$. Je discuterai davantage des différentes variations du modèle log-logistique dans la section 2.3.

$$f(x) = \frac{100}{1 + 10^{s \cdot (\log_{10} c - \log_{10} x)}} \tag{1.3}$$

Modèle Weibull. Le modèle Weibull est principalement utilisé pour décrire des données de survie démographique [18, 19]. Contrairement au modèle log-logistique, ce modèle est asymétrique : le point d'inflexion ne peut pas toujours être associé à l'IC_{50} [12]. Le paramètre c devient donc c'. À titre indicatif, les courbes des figures 1.3a, 1.3c et 1.3d ont un $\log_{10} c = \log_{10} c' = 3.60$. Les valeurs $f(\log_{10} c')$ de ces courbes sont respectivement 50.00 (définition de l'IC_{50}), 10.00 et 90.00.

Le modèle Weibull est souvent décrit selon la convergence de sa courbe, soit "*rapide*" ou "*lente*". Le modèle dit *rapide* (Éq. 1.4) détient un plateau inférieur généralement bien défini. La convergence entre ce plateau et le point d'inflexion se fait rapidement et abruptement. Le reste de la courbe converge plus lentement vers le plateau supérieur qui n'est pas aussi bien défini (Fig. 1.3c p.10).

$$f(x) = a + (b - a) \cdot 10^{-10^{s \cdot (\log_{10} c' - \log_{10} x)}} \tag{1.4}$$

Le modèle dit *lent* (Éq. 1.5) converge quant à lui lentement entre le plateau inférieur et le point d'inflexion, puis rapidement vers le plateau supérieur (Fig. 1.3d p.10) [19, 14].

$$f(x) = a + (b - a) \cdot \left(1 - 10^{-10^{-s \cdot (\log_{10} c' - \log_{10} x)}}\right) \tag{1.5}$$

Il est préférable d'utiliser un des modèles Weibull lorsque les réponses sont globalement faibles (Weibull *rapide*) ou glabalement fortes (Weibull *lente*). Il n'est pas toujours possible de déterminer clairement cette tendance avant d'ajuster le modèle. Dans ces cas d'incertitude, le modèle log-logistique peut s'ajuster de façon satisfaisante à des réponses globalement faibles ou fortes [14]. Cependant, le modèle Weibull *lent* s'ajuste très mal à des réponses globalement faibles et le Weibull *rapide* aux réponses globalement fortes. Dans ces contextes, les estimations des paramètres représentent alors mal les données analysées et peuvent mener à des conclusions biaisées [12].

Modèle Brain & Cousens. Le modèle Brain & Cousens fut présenté pour la première fois en 1988 dans l'optique de modéliser l'hormèse observée pour certaines relations plante-herbicide [20]. L'hormèse se définie comme étant le changement entre la stimulation à basse concentration et l'inhibition à haute concentration [21]. En d'autres termes, l'hormèse est la tendance qu'ont les réponses à faibles concentrations d'être inverses à la tendance globale des réponses à plus hautes concentrations (Fig. 1.3b p.10) [12]. Pour ce type de réponses, il est difficle d'estimer raisonnablement le paramètre a du modèle log-logistique. Comme pour les modèles Weibull, le modèle Brain & Cousens est asymétrique et le point d'inflexion n'est pas représentatif de l'IC_{50}. Il est important de préciser que c' représente l'inflexion de la sigmoïde et non une des inflexions causées par les réponses à basses concentrations [22].

Un nouveau paramètre est introduit dans le modèle Brain & Cousens : h (Éq. 1.6). Celui-ci est une estimation de la pente de la relation linéaire dose-réponse à faibles concentrations.

$$f(x) = a + \frac{b + hx - a}{1 + 10^{s \cdot (\log_{10} c' - \log_{10} x)}} \tag{1.6}$$

Le modèle log-logistique est un cas spécial du modèle Brain & Cousens : lorsqu'il y a absence d'hormèse le paramètre h est égale à 0.00 et l'équation 1.6 devient alors l'équation 1.1 [20].

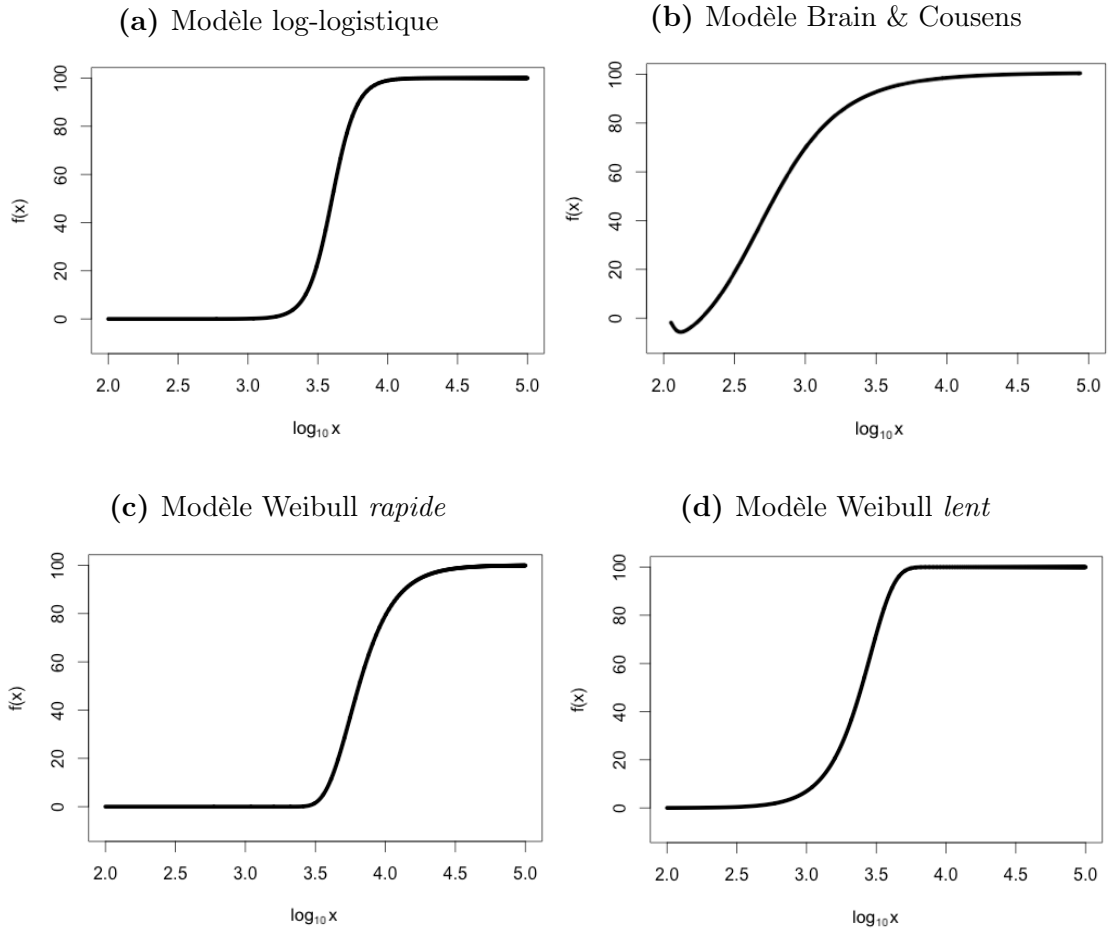


Figure 1.3 – Modèles mathématiques pour la modélisation dose-réponse. Les graphiques illustrent la relation $y = f(x)$ où y sont les réponses et x les concentrations. (a) La sigmoïde est symétrique autour du point d'inflexion. (b) La sigmoïde est asymétrique et modélise les effets anti-inhibitoires à de faibles concentrations. (c)-(d) Les sigmoïdes sont asymétriques et convergent vers le plateau supérieur à différentes vitesses.

1.2 Ajustement de modèle : la régression non-linéaire

Bien qu'il soit pratique de visualiser la tendance globale des données de CHD sous forme de courbe, le but premier de la modélisation est l'extraction de paramètres pour ensuite interpréter ceux-ci selon le contexte expérimental. Par exemple, nous avons vu dans la précédente section que les paramètres c et b peuvent être représentatifs de l'IC_{50} et de la réponse maximale selon le modèle choisi. Il est donc d'intérêt d'ajuster le modèle le plus précisement possible et d'obtenir des estimations valables pour ces paramètres. La présente section traite des approches mathématiques et algorithmiques pour faire un tel ajustement.

Le fait d'ajuster un modèle non-linéaire à un jeu de données est défini par l'appélation "*régression non-linéaire*". Une régression non-linéaire se fait par un processus itératif d'optimisation. L'approche générale vise à optimiser itérativement les paramètres du modèle donné de telle sorte à minimiser une fonction d'erreur. Notons qu'il existe des alternatives à la régression non-linéaire lorsqu'il est question d'ajuster un modèle de dose-réponse. Celles-ci s'inspirent notamment de la modélisation des dynamiques de la croissance cellulaire [23, 24, 25, 26].

Pour bien expliquer l'approche, j'utiliserai l'exemple théorique du modèle $f(x;\theta)$ où θ représente un ensemble de M paramètres associés au modèle et x un ensemble de concentrations. Le modèle devra être ajusté selon le jeu de données théorique D constitué de N observations (x_n, y_n), où x est une concentration, y la réponse telle qu'obtenue par CHD, et n l'index itératif des observations.

1.2.1 La fonction objective

Étant donnés le jeu D, nous savons qu'il existe des θ_j non-représentatifs des données et qu'il existe un $\theta_{\text{réel}}$ qui est le plus représentatif des données. Nous savons aussi qu'il

existe un univers de jeux de données D_j pour $f(x;\theta_{\text{réel}})$ dont notre jeu initial D fait partie. L'objectif est donc d'estimer $\theta \approx \theta_{\text{réel}}$ sans pour autant connaître $\theta_{\text{réel}}$ ou l'univers complet des D_j [27]. Dans cet ordre d'idées, nous désirons alors évaluer la probabilité de D étant donnés θ.

La fonction objective, ou fonction de coût, sert à mesurer cette probabilité et son résultat est indicateur de la façon dont les paramètres devraient être ajustés lors de la prochaine itération. Le calcul de cette probabilité se fait en utilisant la différence entre les valeurs de D et les valeurs prédites par le modèle ajusté, soit les résiduels [27]. Pour une concentration x_n, le residuel r_n est égale à la différence entre y_n et $f(x_n;\theta)$ (Éq. 1.7 & Fig. 1.4 p.13) [28].

$$r_n = y_n - f(x_n;\theta) \tag{1.7}$$

La probabilité de D étant donnés θ se calcule alors en multipliant les probabilités d'obtenir chacune des observations (x_n, y_n) (Éq. 1.8). Lorsque les y_n de D prennent des valeurs continues, une valeur fixe $\pm\Delta y$ leur est ajoutés sans quoi la probabilité serait toujours de zéro.

$$P \propto \prod_{n=1}^{N} \left\{ \exp\left[-\frac{1}{N}\left(\frac{r}{\sigma}\right)^2 \right] \Delta y \right\} \tag{1.8}$$

Nous assumons alors que l'erreur de chaque valeur y_n est indépendante et normalement distribuée. En d'autres mots, y_n est une valeur de la distribution $\mathcal{N}(f(x_n;\theta),\sigma)$. La probabilité de y_n se calcule alors depuis une fonction gaussienne (Éq. 1.8), et la probabilité de D est équivalente à leur produit [27].

Maximiser l'équation 1.8 revient à maximiser la vraisemblance de D étant donné θ. Mathématiquement, cette maximisation est équivalente à minimiser le logarithme négatif

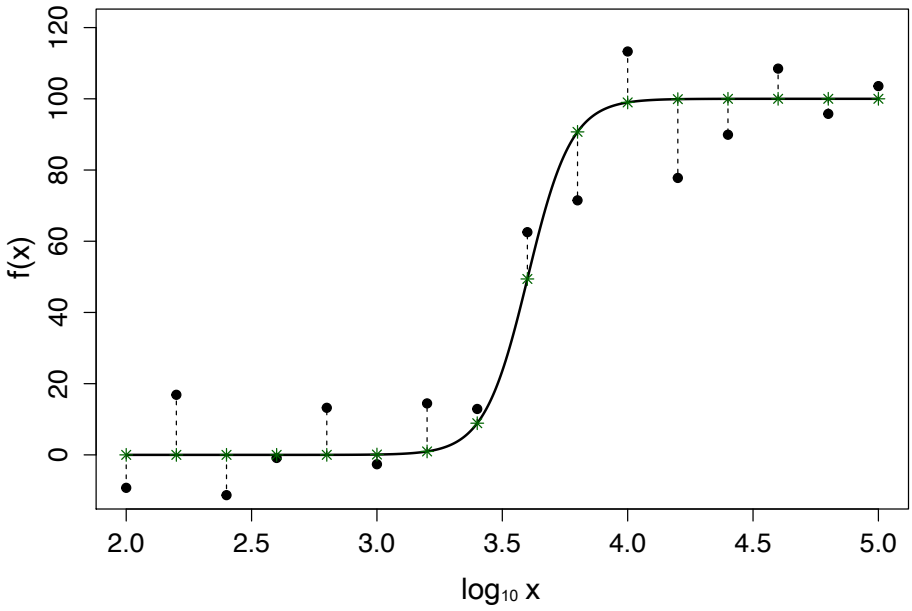


Figure 1.4 – Représentation graphique des résiduels. La sigmoïde noire est obtenue en ajustant le modèle log-logistique aux données synthétiques représentées par les points noirs. Les valeurs $f(x_n;\theta)$ sont représentées par des astérisques verts. Les résiduels sont quant à eux illustrés par les segments verticaux noirs hachés reliant une donnée synthétique à la valeur prédite par le modèle ajusté pour une même valeur de x.

de l'équation 1.8, soit :

$$P = \sum_{n=1}^{N} \left\{ -\frac{r_n^2}{N\sigma^2} \right\} - N \log \Delta y \tag{1.9}$$

Dans ces deux équations, σ et σ^2 représente respectivement l'écart-type et la variance de l'erreur expérimentale. Cette veleur est déterminante de la fonction objective à utiliser. Assumons, dans un premier temps, que l'erreur expérimentale est indépendante et distribuée normalement avec un écart-type constant pour toutes les observations. Puisque N, σ et Δy sont constants, l'équation 1.9 devient celle des moindres carrés (Éq. 1.10). Relaxons notre première supposition et assumons maintenant que l'écart-type de l'erreur expérimentale soit spécifique à chaque observation (σ_n). L'équation 1.9 devient alors celle du chi-carré (Éq. 1.11), où seulement N et Δy sont constants.

$$E(x, \theta) = \sum_{n=1}^{N} r_n^2 \tag{1.10}$$

$$E(x, \theta) = \sum_{n=1}^{N} \left(\frac{r_n}{\sigma_n} \right)^2 \tag{1.11}$$

L'ajustement d'un modèle depuis un jeu de données expérimentales est dicté par la fonction objective $E(x, \theta)$: il est important de bien la choisir. Bien que théoriquement il existe un $\theta_{\text{réel}}$, il faut noter que le θ retourné par notre meilleur ajustement sera toujours une estimation de $\theta_{\text{réel}}$. Cette réalité d'expliquer en deux points : (1) il nous ait impossible d'ajuster le modèle sur tous les D_j possibles et (2) les ressources dont nous disposons pour faire les calculs ont des capacités finies. Cela étant dit, nous sommes aptes à trouver des estimations qui sauront nous satisfaires dans le contexte des analyses menées.

1.2.2 Les algorithmes d'optimisation

La fonction objective, bien qu'essentielle, n'est qu'une composante de la régression non-linéaire. Elle doit être jumelée à un processus d'optimisation itératif pour être effective. Il existe différents algorithmes d'optimisation [29], dont quatre principaux (Table I, p.15). Ces derniers sont semblables dans leur forme, mais divergent dans leur règle d'ajustement. Les différences dans ces règles font varier la convergence des algorithmes (lente ou rapide) ainsi que leur stabilité (stable ou instable). La convergence d'un algorithme est définie comme étant la vitesse à laquelle l'ensemble de ses itérations mène à un point limite. L'algorithme peut être interrompu par un critère d'arrêt. Je définis ce critère comme étant lorsque la différence entre $E(x, \theta_i)$ et $E(x, \theta_{i+1})$ est inférieure à ϵ pour trois itérations consécutives (voir la section 2.2 pour la valeur de ϵ). La stabilité se définit quant à elle comme étant le comportement robuste de l'algorithme à converger vers une solution et ce, pour différents contextes d'initialisation.

Table I – Principaux algorithmes d'optimisation

Algorithme	Règle d'ajustement	Convergence	Complexité de calcul
Descente de gradient	$\theta_{i+1} = \theta_i - \alpha \cdot g_i$	Stable, lente	Gradient
Newton	$\theta_{i+1} = \theta_i - H_i^{-1} \cdot g_i$	Instable, rapide	Gradient, Hessienne
Gauss-Newton	$\theta_{i+1} = \theta_i - \left(J_i^T J_i\right)^{-1} \cdot J_i^T r_i$	Instable, rapide	Jacobienne
Levenberg-Marquardt	$\theta_{i+1} = \theta_i - \left(J_i^T J_i + \mu I\right)^{-1} \cdot J_i^T r_i$	Stable, rapide	Jacobienne

Descente de gradient. L'algorithme de la descente de gradient, aussi connu sous le nom de rétropropagation du gradient, fut proposé dans les années 1970. Il fut réellement mis de l'avant en 1986 avec la publication de l'article de Rumelhart [31]. La technique présentée est alors plus rapide et flexible que les procédures standard de l'époque. Depuis, plusieurs améliorations ont été faites. Malgré cela et bien qu'il soit encore largement utilisé aujourd'hui, l'algorithme converge lentement.

Pour chaque itération i allant de 0 à I, les paramètres θ_i sont ajustés selon une constante d'apprentissage α fixe et le gradient g de la fonction objective (Éq. 1.12).

$$\theta_{i+1} = \theta_i - \alpha g_i \tag{1.12}$$

La constante d'apprentissage α détermine la grandeur des "bonds" lors de l'ajustement. Un large α cause de l'oscillation dans l'ajustement et entraîne le risque d'ajuster les paramètres dans la direction contraire au $\theta_{\text{réel}}$. Pour éviter cela, il est recommandé d'utiliser un petit α. Cela étant dit, l'ajustement se fait alors doucement et à petit pas : pour S itérations, il se pourrait que le modèle n'atteigne pas le θ optimal. La constante d'apprentissage fixe explique la convergence lente et linéaire de la descente de gradient [32, 33].

La descente de gradient est un algorithme dit de premier ordre puisqu'il utilise la dérivée partielle première de la fonction objective, soit le gradient g (Éq. 1.13) [27]. Le vecteur g est de taille M puisque chaque paramètre du modèle ajusté détient son propre gradient [32]. Le calcul du gradient est le seul élément coûteux en ce qui concerne la complexité du calcul [33].

$$g = \frac{\partial E(x, \theta)}{\partial \theta_m} = \begin{bmatrix} \frac{\partial E}{\partial \theta_1} & \frac{\partial E}{\partial \theta_2} & \dots & \frac{\partial E}{\partial \theta_M} \end{bmatrix}^T \tag{1.13}$$

La sélection des méta-paramètres α et S affecte grandement la vitesse de convergence : leur valeur doit être spécifique aux données analysées. De plus, pour un S précis, il est préférable d'initier les paramètres du modèle avec un θ_0 probable plutôt qu'aléatoire (nous aborderons ce sujet plus en détail dans le Chapitre **??**). Dans ce deuxième cas, l'algorithme pourrait ne pas avoir assez de I itérations pour converger vers le θ optimal [32].

Newton. L'algorithme Newton fait référence à la méthode du même nom utilisée en calcul différentiel [34]. Contrairement à la descente de gradient, les bonds entre les ajustements ne sont pas fixes, augmentant ainsi la vitesse de convergence de l'algorithme [33].

Pour chaque itération i, les éléments de θ_i sont ajustés selon le produit de l'inverse de la hessienne H et du gradient g de la fonction objective (Éq. 1.14).

$$\theta_{i+1} = \theta_i - H_i^{-1} g_i \tag{1.14}$$

La hessienne est la matrice carré $M \times M$ des dérivées partielles secondes de la fonction objective (Éq. 1.15) [33].

$$H = \begin{bmatrix} \frac{\partial^2 E}{\partial \theta_1^2} & \frac{\partial^2 E}{\partial \theta_1 \theta_2} & \cdots & \frac{\partial^2 E}{\partial \theta_1 \theta_M} \\ \frac{\partial^2 E}{\partial \theta_2 \theta_1} & \frac{\partial^2 E}{\partial \theta_2^2} & \cdots & \frac{\partial^2 E}{\partial \theta_2 \theta_M} \\ & \cdots & & \\ \frac{\partial^2 E}{\partial \theta_M \theta_1} & \frac{\partial^2 E}{\partial \theta_M \theta_2} & \cdots & \frac{\partial^2 E}{\partial \theta_M^2} \end{bmatrix} \tag{1.15}$$

L'algorithme Newton assume que les composantes du vecteur g sont des fonctions $F(\theta_1, ..., \theta_M)$ où les θ_m sont linéairement indépendants (Éq. 1.16). Les $\Delta\theta_m$ représentent la différence entre les valeurs de θ_{i+1} et θ_i.

$$\begin{aligned} g_m &= F_m(\theta_1, ..., \theta_M) \\ &\approx g_{m,0} + \frac{\partial g_1}{\partial \theta_1}\Delta\theta_1 + ... + \frac{\partial g_M}{\partial \theta_M}\Delta\theta_M \\ &\approx g_{m,0} + \frac{\partial g^2 E}{\partial \theta_m \partial \theta_1}\Delta\theta_1 + ... + \frac{\partial g^2 E}{\partial \theta_m \partial \theta_M}\Delta\theta_M \end{aligned} \tag{1.16}$$

Mathématiquement, lorsque le gradient d'une fonction est égale à 0 cela est indicateur d'un point critique [32]. Par exemple, si $\frac{\partial f(x)}{\partial x} = 0$ alors x est soit un minima ou un maxima de la fonction $f(x)$. Dans ce sens, puisque l'algorithme Newton cherche à minimiser la fonction objective, les composantes du vecteur g doivent toutes être égales à 0 [35]. L'équation 1.16 devient alors :

$$
\begin{aligned}
0 &\approx g_{m,0} + \frac{\partial g^2 E}{\partial \theta_m \partial \theta_1}\Delta\theta_1 + ... + \frac{\partial g^2 E}{\partial \theta_m \partial \theta_M}\Delta\theta_M \\
-g_{m,0} = -\frac{\partial E}{\partial \theta_m} &\approx \frac{\partial g^2 E}{\partial \theta_m \partial \theta_1}\Delta\theta_1 + ... + \frac{\partial g^2 E}{\partial \theta_m \partial \theta_M}\Delta\theta_M
\end{aligned}
\tag{1.17}
$$

L'équation 1.17 s'applique au M composantes de θ et peut donc être écrite sous forme matricielle (Éq. 1.18).

$$
\begin{bmatrix} -g_1 \\ ... \\ -g_M \end{bmatrix} = \begin{bmatrix} -\frac{\partial E}{\partial \theta_1} \\ ... \\ -\frac{\partial E}{\partial \theta_M} \end{bmatrix} \approx \begin{bmatrix} \frac{\partial^2 E}{\partial \theta_1^2} & ... & \frac{\partial^2 E}{\partial \theta_1 \partial \theta_M} \\ & ... & \\ \frac{\partial^2 E}{\partial \theta_M \partial \theta_1} & ... & \frac{\partial^2 E}{\partial \theta_M^2} \end{bmatrix} \times \begin{bmatrix} \Delta\theta_1 \\ ... \\ \Delta\theta_M \end{bmatrix} \tag{1.18}
$$

La matrice $M \times M$ de l'équation 1.18 représente la hessienne telle que définie par l'équation 1.15. Un ré-arrangement de l'équation 1.18 produit la règle d'ajustement telle que présentée par l'équation 1.14 :

$$
\begin{aligned}
-g &= H\Delta\theta \\
\Delta\theta &= -H^{-1}g \\
\theta_{i+1} &= \theta_i - H^{-1}g
\end{aligned}
\tag{1.19}
$$

L'algorithme Newton est plus rapide que la descente de gradient. L'inverse de la hessienne à l'avantage d'ajuster la grandeur des bonds d'ajustement selon la valeur de la fonction objective : pour une valeur se rapprochant d'un point critique (pente abrupte) les bonds seront plus petits, tandis que pour une valeur loin ou s'éloignant d'un point critique (pente douce) les bonds seront plus grands [33]. Cela étant dit, l'algorithme Newton n'est pas garantit de converger vers un θ optimal puisqu'il converge le plus rapidement

possible vers un point critique de la fonction objective, sans faire la distinction entre un minima (souhaité) et un maxima (non souhaité). La convergence de l'algorithme est aussi dépendante du θ initial. Si celui-ci est trop différent du θ optimal, l'alogirthme pourrait ne pas converger. De plus, le calcul de la hessienne augmente la complexité du calcul [32].

Gauss-Newton. L'algorithme Gauss-Newton est une alternative à l'algorithme Newton qui fut proposé pour réduire la complexité du calcul d'ajustement [32]. En effet, les calculs du gradient et de la hessienne sont remplacés par celui de la jacobienne. Cette matrice $N \times M$ représente les dérivés premières partielles des composantes de la fonction de coût (Éq. 1.20), soit les résiduels [27].

$$J = \begin{bmatrix} \frac{\partial r_1}{\partial \theta_1} & \frac{\partial r_1}{\partial \theta_2} & \cdots & \frac{\partial r_1}{\partial \theta_M} \\ \frac{\partial r_2}{\partial \theta_1} & \frac{\partial r_2}{\partial \theta_2} & \cdots & \frac{\partial r_2}{\partial \theta_M} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial r_N}{\partial \theta_1} & \frac{\partial r_N}{\partial \theta_2} & \cdots & \frac{\partial r_N}{\partial \theta_M} \end{bmatrix} \tag{1.20}$$

Les paramètres sont alors ajustés selon l'équation 1.21 où $J^T J$ est une estimation de la hessienne tandis que $J^T r$ est égale au gradient. Pour une itération i, l'ajustement $i+1$ se fait tel que :

$$\theta_{i+1} = \theta_i - (J_i^T J_i)^{-1} J_i^T r_i \tag{1.21}$$

Incorporer l'équation de la fonction objective (1.10) à celle du gradient (1.13) démontre l'égalité $g = J^T r$:

$$g = \frac{\partial E}{\partial \theta_m} = \frac{\partial \sum_{n=1}^{N} r_n^2}{\partial \theta_m} = \sum \frac{\partial r_n}{\partial \theta_m} r_n = J^T r \tag{1.22}$$

Pour ce qui est de l'estimation de la hessienne, il suffit d'insérer la fonction objective

(Éq. 1.10) dans l'une des composantes de la hessienne :

$$H_{m,m} = \frac{\partial^2 E}{\partial\theta_m \partial\theta_m} = \frac{\partial(\partial \sum_{n=1}^{N} r_n^2)}{\partial\theta_m \partial\theta_m} = \sum_{n=1}^{N} \frac{\partial^2 r_n}{\partial\theta_m \partial\theta_m} r_n + \sum_{n=1}^{N} \frac{\partial r_n}{\partial\theta_m} \frac{\partial r_n}{\partial\theta_m} \approx J^T J \tag{1.23}$$

L'approximation ci-haut se vaut puisque nous assumons que $\sum_{n=1}^{N} \frac{\partial^2 r_n}{\partial\theta_m \partial\theta_m} r_n$ tend vers 0. Cela va de soit puisque l'algorithme Newton assume que les composantes de g soient égales à 0 (Éq. 1.17) [35].

Bien que l'agorithme Gauss-Newton soit plus efficace en terme de calcul et donc plus rapide que l'algorithme Newton, sa convergence vers le θ optimal n'est toujours pas garantie. Pour certaines valeurs, l'estimation de la hessienne (Éq. 1.23) est singulière et la composante $(J^T J)^{-1}$ de l'équation 1.21 ne peut être calculée [33].

Levenberg-Marquardt. Développé dans un premier temps par Levenberg [36] puis par Marquardt [37], l'algorithme Levenberg-Marquardt minimise la fonction objective en alternant entre la descente de gradient et l'algorithme Gauss-Newton. Cette approche permet de converger vers le θ optimal rapidement. Elle est adéquate pour des problèmes de petite et moyenne taille, et est aujourd'hui considérée comme l'approche standard pour les problèmes de minimisation des carrés (Éq. 1.10 et 1.11) [33].

Tout comme pour l'algorithme Gauss-Newton, seule la jacobienne est calculée. L'approximation de la hessienne est cependant différente : la composante μI lui ait ajoutée, où μ est le coefficient de l'approximation hessienne et I la matrice identité. Les paramètres θ_i sont alors ajustés selon l'équation 1.24.

$$\theta_{i+1} = \theta_i - (J_i^T J_i + \mu I)^{-1} J_i^T r_i \tag{1.24}$$

La convergence stable de l'algorithme Levenberg-Marquardt s'explique par la nouvelle approximation de la hessienne. Celle-ci garantit que les éléments de la diagonale seront toujours supérieurs à 0 et par le fait même que la matrice soit inversible. La vitesse de convergence s'explique quant à elle par l'alternance entre la descente de gradient et l'algorithme Gauss-Netwon. Cette alternance est déterminée par la valeur du coefficient de l'approximation hessienne. Pour un μ très petit, l'équation 1.24 se rapproche de l'équation 1.21, et Gauss-Newton est utilisé ; lorsque μ est plutôt large, la règle d'ajustement se rapproche plutôt de celle de la descente de gradient (Éq. 1.12).

Il y a alternance entre les approches puisque la valeur de μ fluctue selon la relation entre $E(x, \theta_i)$ et $E(x, \theta_{i+1})$, tel que démontré par l'algorithme 1. Dans un premier temps, θ_0 et μ_0 sont initialisés, tout comme le nombre d'itération S et le nombre d'essais T. Pour une itération i, θ_i est ajusté selon l'équation 1.24 pour devenir θ_{i+1}. La fonction objective est évaluée pour θ_i et θ_{i+1}. Si $E(x, \theta_{i+1}) > E(x, \theta_i)$, alors l'ajustement θ_{i+1} est rejeté, μ est mutiplié par un facteur λ et un nouveau θ_{i+1} est calculé. Ce processus est répété pour un maximum de T essais, après lesquels l'ajustement θ_{i+1} est accepté peu importe la valeur de $E(x, \theta_{i+1})$. Si $E(x, \theta_{i+1}) \leq E(x, \theta_i)$, alors l'ajustement θ_{i+1} est accepté et μ est diviser par λ. L'index i ne peut être incrémenté que lorsqu'un ajustement θ_{i+1} est accepté.

L'algorithme Levenberg-Marquardt a la convergence la plus rapide et la plus efficace de tous les algorithmes d'optimisation présentés ci-haut.

1.3 Outils de calcul

Les sections précédentes présentaient les deux composantes clés pour l'analyse des résultats de CHD de type dose-réponse : (1) les modèles pour la modélisation des données et l'extraction des métriques d'intérêt, et (2) les approches algorithmiques pour l'ajustement de tels modèles à des données. La présente section présentera les trois principaux outils représentatifs pour l'analyse des données de CHD, bien qu'il existe divers outils

Entrées : D
Sorties : θ

```
1  def ajustement(θ_i, μ, t):
2      θ_{i+1} = θ_i − (J_i^T J_i + μI)^{-1} J_i r_i
3      si E(x, θ_{i+1}) > E(x, θ_i) alors
4          μ = μ × λ
5          si t > T alors
6              retourner θ_{i+1}, μ
7          fin
8          t = t + 1
9          retourner ajustement(θ_i, μ, t)
10     sinon
11         μ = μ ÷ λ
12         retourner θ_{i+1}, μ
13     fin
14 initialisation θ, μ, T, S
15 pour i = 0 à S faire
16     t = 0
17     θ_i = θ
18     θ, μ = ajustement(θ_i, μ, t)
19 fin
20 retourner θ
```

Algorithme 1 : Levenberg-Marquardt

[17, 38, 39, 40]. Ceux-ci peuvent être selon moi classifiés en deux groupes : (1) ceux qui sont simples d'usage (de l'anglais *user-friendly*) et (2) ceux qui nécessitent une certaine expertise.

Je définis ce premier groupe par des outils qui sont à la portée de chercheurs de différents domaines (biologie, chimie, etc.). Leur usage est simple et peut s'apprendre relativement facilement. De plus, ils sont souvent présentés sous la forme d'une interface graphique. Un example d'un tel outil est GraphPad de Prism [www.graphpad.com]. Bien qu'il propose des options avancées qui requièrent un peu plus de connaissances, son usage de base est simple et direct.

Le deuxième groupe se définit alors plutôt par des outils nécessitant une connaissance plus accrue du processus d'analyse. ActivityBase et SARview d'IDBS [www.idbs.com], l'environnement R [www.r-project.org] ainsi que les multiples librairies Python en sont des exemples. Ces-derniers nécessites souvent une ou plusieurs séances de formation pour comprendre leur fonctionnement et la mécanique de l'outil. Ils requièrent aussi souvent une compréhension de l'informatique ainsi qu'une maîtrise de la programmation.

Il est à noter que ces catégories ont été définies subjectivement selon un consensus approximatif que j'ai observé lors des mes interactions avec divers collaborateurs. De plus, bien que la grande majorité des outils du premier groupe prennent la forme d'une interface graphique, cette caractéristique n'est pas unique à ce groupe. Effectivement, des outils du deuxième groupe tel qu'activityBase ont la forme d'une interface.

Ces deux catégories d'outils retournent sensiblement les mêmes informations, soit une courbe représentant l'ajustement, les estimations finales des paramètres, et parfois un intervalle de confiance sur l'ajustement et/ou les paramètres. Il est présentement difficile, voire impossible dans certains cas, de faire une comparaison de deux ajustements en utilisant les outils listés plus haut. Généralement, une telle comparaison se fait qualitivement en observant les courbes de l'ajustement, et un tant soit quantitivement en comparant les valeurs estimées des IC_{50}. La majorité de ces outils ne sont pas automatisés et un travail de traitement de données est requis de la part du chercheur. De plus, ces outils peuvent être très coûteux monétairement et en temps d'analyse. Ils requièrent plusieurs étapes au processus d'analyse qui sont bien souvent faites manuellement par le chercheur (exemple : extraction des données, normalization, etc.). Cela étant dit, tout travail manuel demande un effort supplémentaire de gestion pour assurer sa traçabilité et sa reproductibilité. Le processus d'analyse complet des outils disponibles est souvent *caché* et le jeu de données n'est pas toujours exploité à son plein potentiel informationnel.

Le présent travail vise à mettre sur pied une méthodologie et un processus pour l'analyse de données de CHD. Le processe se voudra automatisé, flexible et statistique. À notre connaissance, aucun des outils présentement utilisés ne combine ces trois caractéristiques. Le chapitre qui suit présentera en détail la méthodologie ainsi que le processus qui lui est associé.

Chapitre 2

Méthodologie pour l'analyse des données de criblage à haut débit

Le présent travail fut entrepris dans le contexte de plusieurs collaborations avec divers biologistes et chimistes. Tous utilisent le criblage à haut débit (CHD) et cherchent à obtenir des estimations pour des paramètres du modèle log-logistique. De ces collaborateurs, aucun n'a mentionné utiliser des outils de programmation tels que des librairies R et Python pour faire leurs analyses. Celles-ci se font avec GraphPad ou avec ActivityBase/SARview.

L'avantage principal de GraphPad réside dans le calcul d'intervalles de confiance pour chacun des paramètres estimés. Cependant, un gros désavantage de cet outil d'analyse est la quantité de travail manuel qui doit être fait. Chaque donnée utilisée pour l'ajustement doit être entrée manuellement par le chercheur dans l'interface. Dans le contexte du CHD, l'analyse d'une expérience testant 10 000 composés devient très lourde en plus d'être coûteuse en terme de temps et propice à des erreurs de manipulation de données. Bien qu'il soit possible de faire de telles analyses avec GraphPad, cette approche n'est pas optimale.

ActivityBase/SARview est quant à lui plus optimal en ce qui a trait à la gestion des données. Effectivement, l'outil analyse les données brutes après leur mesure. Il est

alors beaucoup simple d'analyser un grand nombre de composés. Cependant, le protocoles d'analyse doit être déterminé avant même que le criblage soit fait. Une fois l'analyse faite, il est difficile de retourner aux données brutes et faire une nouvelle analyse. Pour ce faire, il faudrait refaire une expérience de criblage ayant des conditions semblables.

Bien que les deux outils principalement utilisés présentent des avantages, leurs inconvénients peuvent dérouter une expérience et son analyse. De plus, aucun de ces outils ne propose une comparaison statistique de deux ajustements. Il est vrai qu'il existe une option *Compare* dans GraphPad, mais celle-ci tente plutôt de déterminer lequel des modèles log-logistique (3 ou 4 paramètres) est le plus approprié pour un jeu de données.

L'objectif du présent travail est donc de mettre sur pied une approche automatisée, flexible et statistique pour l'analyse des données de CHD (Table II p.25). L'automatisation du processus facilitera grandement l'analyse qui ne sera pas contrainte pas la quantité de données ni par le nombre d'ajustements à faire. La flexibilité de l'approche élargira les possibilités d'application (voir la section 3.5). Finalement, l'approche statistique permettra d'établir l'exactitude des estimations de modèle log-logistique en plus de données une valeur à la comparaison d'ajustements.

Table II – Outils pour l'analyse de données de CHD

	Approche proposée	**GraphPad**	**ActivityBase/SARview**
Automatisation	✓	✗	✓
Flexibilité	✓	∼	∼
Statistiques	✓	∼	✗

Descriptifs des différentes caractéristiques : présente (✓), partiellement présente (∼), absente (✗)

Pour obtenir le processus d'analyse proposé, cinq aspects ont été étudiés, puis combinés : (1) l'implémentation de la régression non-linéaire, (2) la sélection du modèle permettant une exploitation optimale des données, (3) le calcul d'intervalles de confiance, (4) la comparaison de deux ajustements, et (5) l'analyse dite de *groupe*. Chacune de ces composantes seront présententées et approfondies dans les sections qui suivent.

Différentes approches ont été testées pour chacune des composantes listées ci-haut.

Pour bien cerner leurs avantages et pour bien comprendre leurs effets, des expériences d'ajustement de modèle ont été faites sur des données synthétiques.

2.1 Données synthétiques

Différents protocoles d'analyse ont été testée dans le but de déterminer lequels seraient les plus efficaces dans le contexte d'analyse de données de CHD. La tâche de quantifier et comparer l'efficacité de ces protocoles est loin d'être triviale. Je définis *protocole efficace* comme étant rapide, précis et généraliste. L'utilisation de données synthétiques me permet d'évaluer ces trois critères dans un environnement contrôlé. Ces données sont générées depuis le modèle log-logistique. Des réponses y sont calculées pour un ensemble de concentrations hypothétiques x et pour des valeurs de a, b, c et s données. Un bruit gaussian aléatoire est par la suite ajouté à chacune de ces réponses. Pour simuler des données bruitées, la distribution gaussienne à un écart-type égale à 15.0. Dans l'optique de reproduire au mieux des contextes expérimentaux, j'ai aussi considéré un jeu de données avec quelques données aberrantes (de l'anglais *outliers*). Le bruit ajouté provient alors d'un modèle de mélanges gaussiens (de l'anglais *Gaussian mixture*) pour $\sigma_1 = 5.0$ et $\sigma_2 = 20.0$. Chaque donée synthétique générée a 5% de chance d'être aberrante, c'est-à-dire que le bruit ajouté proviennent de la distribution ayant un écart-type de 20.0.Les réponses synthétiques représentent un taux (%) qui augmente en importance avec les concentrations. Un exemple d'un tel taux est l'inhibition de la croissance cellulaire en présence d'un composé chimique.

Les expériences d'ajustement faites sur des données synthétiques ont aussi été par la suite entreprises sur des données expérimentales. De la sorte, les conclusions faites lors du processus de développement et dans un environnement contrôlé, peuvent être confirmées (ou infirmées) dans un contexte expérimental. Le Chapitre 3 traite des expériences faites dans ce contexte.

2.2 Implémentation de la régression non-linéaire

Après avoir implémenté et testé les différents alogirthmes présentés dans la section 1.2, j'ai décidé d'utilisé l'algorithme de Levenberg-Marquardt pour sa convergence stable et rapide, ainsi que pour la simplicitié du le calcul des ajustements (Table I p.15). De plus, l'algorithme a l'avantage d'être applicable à plusieurs jeux de données différents, contrairement à la descente de gradient dont la constante d'apprentissage α doit souvent être spécifique au jeu analysé. Un ajustement préliminaire de cette constante est donc nécessaire, ce qui n'est pas optimal dans le contexte du présent travail. Pour faciliter l'utilisation d'équations et de divers concepts mathématiques, la régression est implémentée en Python [www.python.org] à l'aide de la librairie Theano [www.deeplearning.net/software/theano/]. La matrice jacobienne ainsi que l'inversement matriciel sont obtenus grâce à des fonctions intégrées.

Processus d'ajustement. L'analyse se fait sur un ensemble de N observations pairées de type (x_n, y_n) où x est une concentration et y la réponse associée. Une fois l'ajustement fait, une estimation pour chaque paramètre est retournée ainsi que l'erreur moyenne quadratique (EMQ, de l'anglais *root mean square error* ou *RMSE*) (Éq. 2.1). L'EMQ est un indicateur de la précision de l'ajustement. Elle représente l'écart moyen entre la courbe prédite et les observations. Ses unités sont les mêmes que pour les y. Puisque l'EMQ est plus simple à interpréter que les moindres carrés et que nous souhaitons minimier ces deux valeurs, l'algorithme d'optimisation a pour fonction objective l'équation 2.1.

$$EMQ = \sqrt{\frac{\sum_{n=1}^{N} r_n^2}{N}} \tag{2.1}$$

Données synthétiques. Deux types de jeux de données synthétiques ont été utilisés pour obtenir les résultats présentés dans cette section : (1) avec données bruitées ($\sigma = 15$) et (2) avec données aberrantes ($\sigma_1 = 5$, $\sigma_2 = 20$) (voir section 2.1 pour plus de détails sur la génération des données synthétiques). Pour ces deux types, chaque jeu est composé

de 50 observations réparties en 5 réplicats (nombre de réponses par concentration) pour 10 concentrations hypothétiques. Les paramètres utilisés pour générer les données sont les mêmes pour les deux types de jeux soit $a = 0.00$, $b = 100.00$, $\log_{10} c = 2.40$ et $s = 1.00$.

Initialisation des méta-paramètres. L'algorithme Levenberg-Marquardt est reconnu pour l'efficacité de sa convergence lorsque appliqué à des problèmes de petite et moyenne tailles. Il est généralement capable de converger vers les paramètres optimaux en 50 itérations ou moins [33]. Compte tenu la taille relativement petite des problèmes analysés dans le présent travail (4 paramètres à ajuster selon N données), plusieurs expériences confirment que 50 itérations sont amplement suffisantes pour atteindre des paramètres optimaux. De plus, une initiation adéquate des différentes composantes de l'algorithme aide à minimiser le nombre d'itérations et donc d'optimiser le temps de converge. Le temps de convergence, en terme de nombre d'itérations, est affecté par les valeurs des méta-paramètres. Ces derniers sont l'initiation du coefficient de l'approximation hessienne μ et le choix du facteur d'ajustement λ. Dans le but d'augmenter l'efficacité des ajustements, des expériences ont été menées pour mieux comprendre les effets des méta-paramètres. Les paramètres du modèle à ajuster étaient quant à eux initialisés arbitrairement.

Au total, douze combinaisons de méta-paramètres ont été testées. Le coefficient de l'approximation hessienne initial (μ_0) varie entre 0.001, 1.0 ou 100.0. Pour ce qui est du facteur d'ajustement λ (voir l'algorithme 1), j'ai décidé d'implémenter l'algorithme avec λ_1 et λ_2 tel que $\mu = \begin{cases} \mu \times \lambda_1 & \text{si } E(x, \theta_{i+1}) > E(x, \theta_i) \\ \mu \div \lambda_2 & \text{sinon} \end{cases}$. La valeur du facteur d'ajustement est donc spécifique au résultat de la fonction objective étant donné θ_{i+1}. Cela a pour effet de faire fluctuer μ plus précisément et à la capacité d'améliorer la vitesse de convergence de l'algorithme. Les facteurs d'ajustement λ_1 et λ_2 peuvent être soit grand (100.0), soit petit (10.0). Les différentes combinaisons de méta-paramètres ont chacune été testées sur 1 000 jeux de données synthétiques bruitées ($\sigma = 15$) tels que décrits plus haut. L'efficacité des approches est représentée par la moyenne du nombre d'itérations requises pour que l'ajustement converge vers des paramètres optimaux (Fig.2.1 p.29).Je dis d'un ajustement

qu'il a convergé lorsque $|E(x, \theta_i) - E(x, \theta_i)| < \epsilon$ (où $\epsilon < 1 \times 10^{-6}$) pour trois itérations consécutives.

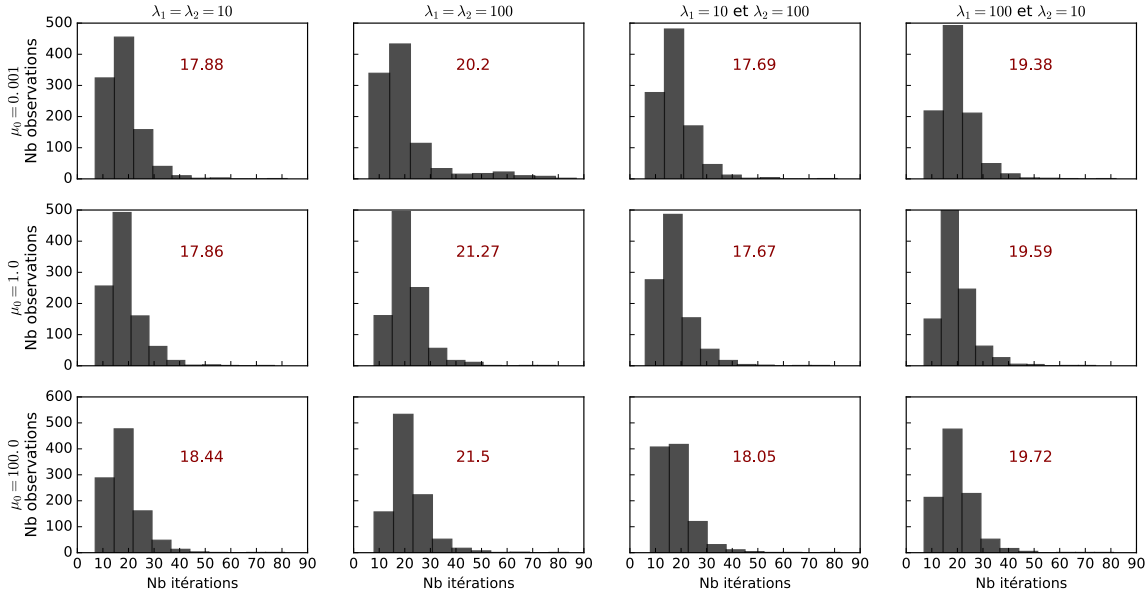


FIGURE 2.1 – EFFETS DES MÉTA-PARAMÈTRES SUR LA VITESSE DE CONVERGENCE. La vitesse de convergence est décrite par le nombre d'itérations nécessaires pour converger. La moyenne pour 1 000 jeux de données synthétiques bruitées ($\sigma = 15$) est inscrite en rouge pour chaque ensemble de méta-paramètres. Les valeurs de μ_0 testées varient entre 0.001, 1.0 et 100.0, tandis que les valeurs de λ_1 et λ_2 varient entre 10.0 et 100.0. Les graphiques sur une même rangée ont le même μ_0 et ceux dans une même colonne partagent les mêmes λ_1 et λ_2.

Étonnamment, lorsque l'on compare les expériences décrites plus haut, les nombres d'itération moyens sont très semblables. La moyenne la plus élevée est de 21.5 itérations ($\mu_0 = 100.0, \lambda_1 = \lambda_2 = 100.0$) et la moins élevée de 17.67 itérations ($\mu_0 = 1.0, \lambda_1 = 10, \lambda_2 = 100.0$). Cette petite différence d'à peine quatre itérations démontre bien la robustesse de l'algorithme Levenberg-Marquardt. De façon générale, il semble plus efficace d'initier μ_0 à une petite valeur. Il semble aussi plus efficace d'utiliser un petit λ_1. La valeur prise par λ_2 semble quant elle avoir moins d'effet sur le nombre moyen d'itérations nécessaires pour converger

La force de Levenberg-Marquardt réside dans l'aternance entre les algorithmes de descente de gradient et Gauss-Newton.Lorsque μ est très petit, l'algorithme tend vers la descente de gradient où la constante d'apprentissage α est égale à $\frac{1}{\mu}$. L'ajustement des paramètres est alors plus important lorsque μ est petit, augmentant ainsi la vitesse de convergence. Plus nous nous approchons des valeurs optimales, plus la valeur de μ est divisiée par λ_2. Lorsque nous nous éloignons des valeurs optimales, μ est multiplié par λ_1 : si nous augmentons trop la valeur de μ, le processus d'ajustement ralentit. Cela explique pourquoi les ajustements sont plus lents lorsque $\mu_0 = 100.0$ et pourquoi les ajustements ayant $\lambda_1 = 100.0$ sont les plus lents, pour tous μ_0 confondus.

Bien que l'algorithme converge vers les valeurs optimales dans tous les cas et ce rapidement, il semble tout de même préférable d'initier μ_0 à une petite valeur et d'utiliser $\lambda_1 = 10.0$.

Initialisation des paramètres à ajuster. La sélection de bons méta-paramètres optimise le temps de convergence de l'algorithme Levenberg-Marquardt, tel que démontré par la figure 2.1. Un autre aspect du processus d'ajustement à considérer est l'initiation des paramètres à ajuster. Les résultats obtenus plus haut sont représentatifs d'une initiation arbitraire où $\theta_0 = [a_0, b_0, \log_{10} c_0, s_0] = [58.00, 24.00, 1.90, 1.70]$[1]. Selon moi, il serait possible d'améliorer le temps de convergence en initiant θ_0 à des valeurs se rapprochant des paramètres optimaux. Cependant, comment pouvons-nous anticiper ces valeurs optimales ? Les données utilisées pour l'ajustement sont indcatrices des futures paramètres ajustés et devraient donc être utilisées lors de l'initiation. J'ai étudié les effets de deux approches d'initialisation se basant sur les données :

- *Initialisation par percentile.* Les valeurs de a_0 et b_0 sont égales aux 5^e et 95^e percentiles, respectivement,des réponses y utilisées pour l'ajustement. La valeur de c_0 est égale au 50^e percentile des concentrations x, et s_0 est arbitrairement initié à 1.00.

[1]Les paramètres initiaux ont été sélectionnés aléatoirement depuis une distribution normale, soit $\mathcal{N}(50, 20)$ pour les paramètres a_0 et b_0, $\mathcal{N}(2, 1.5)$ pour le paramètre $\log_{10} c_0$, et $\mathcal{N}(1, 0.5)$ pour le paramètre s_0

- *Initialisation selon les extrêmes.* Les valeurs de a_0 et b_0 sont égales aux réponses minimale et maximale observées, respectivement, pour toutes concentrations confondues. La valeur de c_0 est égale à $C_{min} + \frac{C_{max} - C_{min}}{2}$ où C_{min} et C_{min} sont les concentrations minimale et maximale testées. La valeur de s_0 est arbitrairement initiée à 1.00.

Ces deux approches d'initiation sont testées sur 1 000 jeux de données synthétiques bruitées ($\sigma = 15$). Les ajustements les plus rapides étaient obtenus lorsque $\mu_0 = 1.0$, $\lambda_1 = 10$ et $\lambda_2 = 100$. Lorsque l'on compare le nombre d'itérations moyen pour converger des trois approches d'initiation, les approches par percentile et selon les extrêmes semblent être plus efficaces que la simple initiation arbitraire (Fig.2.2, p.31). Bien que l'approche par percentile ait une moyenne inférieure à celle des extrêmes, le gain est minime (à peine une itération).

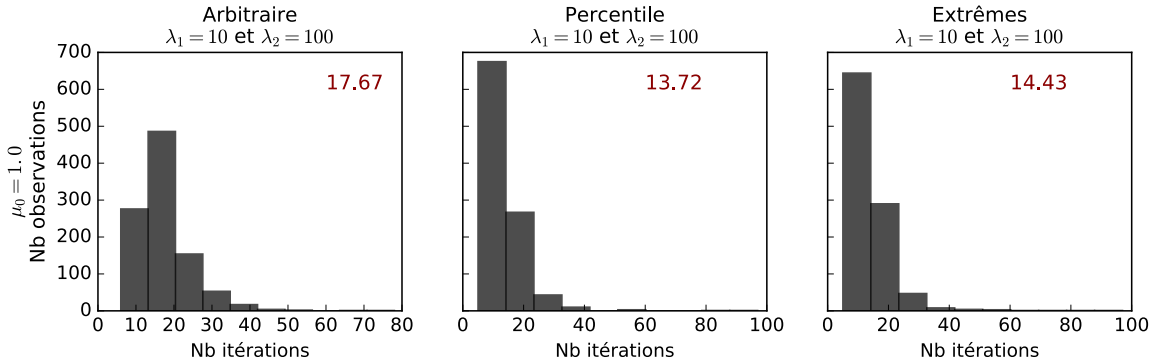


Figure 2.2 – Effets de l'initialisation des paramètres à ajuster sur la vitesse de convergence. La vitesse de convergence est décrite par le nombre d'itérations nécessaires pour converger. La moyenne pour 1 000 jeux de données synthétiques bruitées ($\sigma = 15$) est inscrite en rouge pour chaque approche d'initiation. Les ajustement ont tous pour méta-paramètre $\mu_0 = 1.0$, $\lambda_1 = 10$ et $\lambda_2 = 100$.

Tout comme pour les méta-paramètres, le choix des paramètres initiaux affecte le temps de convergence de l'algorithme. Encore une fois, la différence entre les approches n'est pas très grande. Cependant, si l'on combine les effets des méta-paramètres et de l'initiation des paramètres, le meilleur scénario d'ajustement converge en moyenne en 13.72 itérations

(Fig.2.2 p.31) comparé à 21.5 itérations pour le pire scénario (Fig.2.1 p.29). Sur une station standard (i7-3770 @ 3.4 GHz), une itération prend 0.13 secondes pour les jeux de données synthétiques analysés ci-haut. Le meilleur scénario détient alors un gain de 1.01 secondes sur le pire scénario. Bien que ce gain puisse paraître négligeable, il est hautement bénéfique lors du calcul des intervalles de confiance. Nous verrons dans la section 2.4 que les intervalles sont issues d'un processus répétitif. L'implémentation du meilleur scénario génère alors un gain d'au moins 17 minutes[2], ce qui n'est pas négligeable.

Robustesse de l'algorithme Levenberg-Marquardt. L'algorithme converge rapidement et correctement même lorsque les données sont bruitées. Pour bien confirmer la robustesse de l'algorithme Levenberg-Marquardt, je l'ai appliqué à un jeu de données synthétiques avec des valeurs aberrantes. Dans le contexte de l'analyse de données de CHD, nous travaillons avec des données expérimentales et la présence de données aberrantes est probable. Plusieurs facteurs expérimentaux peuvent générer de telles données. Il est donc essentiel de s'assurer que le processus d'ajustement sélectionné soit efficace en présence de ces données aberrantes.

Ayant conclu que l'initiation arbitraire n'était pas la plus efficace, seules les approches par percentile et selon les extrêmes sont testées. Comme pour toutes les expériences précédentes, l'ajustement semble plus efficace lorsque $\lambda_1 = 10.0$ et $\lambda_2 = 100.0$. Il semble aussi encore préférable d'initier μ_0 à une petite valeur. L'initiation des paramètres par percentile est seulement un peu plus rapide que celle selon les extrêmes (Fig.2.3 p.33). Cette petite différence peut être expliquée par les cas où les données extrêmes utilisées pour l'initiation sont des données aberrantes. Bien que ces valeurs proviennent du jeu de données utilisés pour l'ajustement, elles ne sont pas représentatives des paramètres optimaux. Il est donc attendu que pour ces cas précis l'ajustement soit un peu plus lent.

Somme toute, l'algorithme Levenberg-Marquardt est assez robuste pour converger vers les paramètres optimaux même lorsqu'il y a présence de données aberrantes. Pour maxi-

[2]Gain calculé pour 1000 ré-échantillonnages bootstrap ou 1000 simulations Monte-Carlo, voir la section 2.4 pour plus de détails sur le calcul des intervalles de confiance.

miser le processus d'ajustement des analyses qui suivront, j'utiliserai $\mu_0 = 1.0$, $\lambda_1 = 10.0$ et $\lambda_2 = 100.0$ comme méta-paramètres et initierai les paramètres à ajuster selon l'approche par percentile. La combinaison de ces approches sera dès lors référée comme étant le *processus d'ajustement par défaut* (PAD).

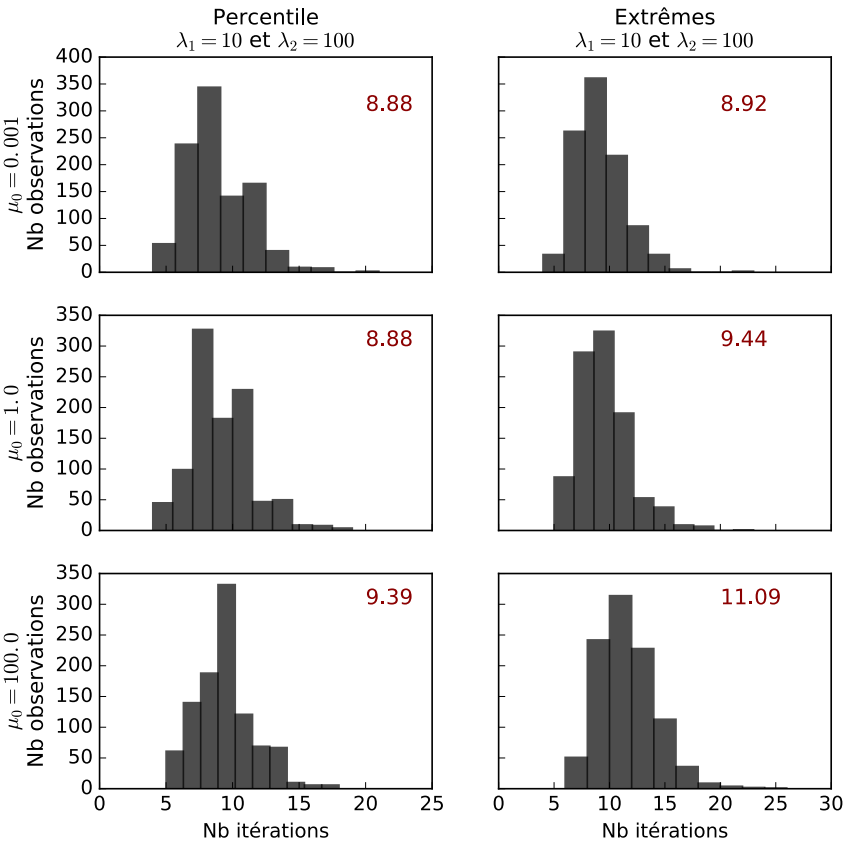


Figure 2.3 – Effets des données aberrantes sur la vitesse de convergence.La vitesse de convergence est décrite par le nombre d'itérations nécessaires pour converger. La moyenne pour 1 000 jeux de données synthétiques ($\sigma_1 = 5, \sigma_2 = 20$) est inscrite en rouge pour les approches d'initiation par percentile et selon les extrêmes. Les ajustements sont effectués pour différents μ_0 et pour $\lambda_1 = 10$ et $\lambda_2 = 100$.

2.3 Exploitation optimale du modèle log-logistique

Tel qu'indiqué dans le titre de la présente section, le modèle mathématique ajusté est le log-logistique. Les deux modèles de Weibull ainsi que le modèle de Brain & Cousens présentés dans le Chapitre 1 doivent être utilisés dans un contexte de données bien précis, sans quoi les estimations de paramètres pourraient être biaisées. Cela étant dit, les données analysées au Chapitre 3 ne correspondent pas à ces contextes précis. Le modèle log-logistique peut quant à lui bien modéliser différents contextes de données (nous verrons dans le Chapitre 3 que le jeu de données comprend plusieurs patients et plusieurs composés). Il est à noter que toutes les analyses et processus décrits dans les chapitres 2 et 3 sont applicables peu importe le modèle sélectionné.

Tel que présenté dans le Chapitre 1, le modèle log-logistique contient quatre paramètres de base (Éq. 1.1). Or, un bon nombre de variantes peuvent être générées en fixant un ou plusieurs de ces paramètres à une valeur constante. Dans le contexte d'analyse de données de CHD, deux variantes du modèle viennent en tête : le modèle à deux paramètres où a et b sont constants (Éq. 1.2), et le modèle à trois pramètres où seulement a est constant (Éq. 1.2). Bien que différents arguments soutiennent l'utilisation de ces variantes, les paramètres estimés ne sont, selon moi, pas représentatifs de la réalité des données. Les conclusions tirées de l'analyse des ajustements de ces variantes risquent d'être erronées et faites en ne considérant qu'une part d'informations. Je tente de démontrer cela en analysant des données synthétiques.

Données synthétiques. Deux jeux de données synthétiques légèrement bruitées ($\sigma = 5$) sont générés pour différentes valeurs du paramètre b. Le jeu de données *noir* a une réponse maximale (b) de 100.00 tandis que celle du jeu de données *vert* est de 50.00. Les autres paramètres sont sensiblement les mêmes pour les deux jeux, soit $b \approx 0.00$, $\log_{10} c \approx 2.40$ et $s \approx 1.00$. Chaque jeu contient 50 observations réparties en 5 réplicats (nombre de réponses par concentration) pour 10 concentrations hypothétiques.

Ajustements des différentes variantes. Les deux variantes du modèle log-logistique ainsi que le modèle même sont ajustés pour chacun des jeux synthétiques (Fig. 2.4 p.37 & Table[3] III p.39). Tel qu'attendu, les estimations de paramètres pour un même jeu de données diffèrent d'une variante à l'autre.

La première variante, soit celle à deux paramètres, s'appuie entièrement sur le contexte biologique expérimental du CHD. Il est tout à fait logique de fixer le paramètre a à 0.00 puisqu'à concentration infiniment petite, un composé devrait générait aucune réponse. Inversement, pour une concentration infiniment grande, un composé devrait générer une réponses maximale (b) de 100.0%. L'ajustement *noir* de la figure 2.4a semble être représentatifs des données et les paramètres estimées sont semblables à ceux utilisés pour générer les données synthétiques. Cependant, l'EMQ de l'ajustement est de 34.07 (Table III p.39). Considérant que les réponses de cet ajustement sont comprises entre 0.00 et 100.00, une erreur de 34.07 est assez large. Pour ce qui est de l'ajustement *vert* de la figure 2.4a, celui-ci n'est évidemment pas représentatif des données et les paramètres estimés confirment bien cela. De plus, l'EMQ de cet ajustement est quasiment le double de celui de l'ajustement *noir*, soit 69.31 (Table III p.39).

La deuxième variante, soit celle à trois paramètres, est une version relaxée de la première variante. Le paramètre a est toujours fixé à 0.00, mais le paramètre b est maintenant ajusté. Il est possible qu'un composé soit incapable de générer une réponse maximale de 100%. D'un point de vue biologique, il est vrai qu'à concentration infiniment grande la réponse observée soit de 100%. Cependant, nous ne pouvons pas affirmer que cette réponse soit entièrement due au composé chimique. Il se pourrait qu'à hautes concentrations de nouveaux facteurs affectant la réponse étudiée entrent en jeu. Penons le contexte de l'étude des effets d'un composé sur l'inhibition de la croissance cellulaire comme exemple. Pour une très grande concentration de composé, il se pourrait que la quantité de nutriments dans le média ne soit plus suffisante compte tenu de l'espace restrain de la réaction. Le

[3]Le nombre de valeurs après la décimal n'est pas représentatif de la précision des estimations de paramètres. La section 2.4 aborde le calcul d'intervalle de confiance qui sont eux représentatifs de la précision des estimations.

taux d'inhibition élevé ne serait alors pas nécessairement causé par les effets du composé. Encore une fois, l'ajustement *noir* semble bien représenter les données (2.4b p.37). Les paramètres estimés sont très semblables à ceux obtenus pour la première variante et à ceux utilisés pour générer les données. Cependant, l'EMQ est maintenant beaucoup plus petite (5.84) et se rapproche de la valeur de l'écart-type du bruit ajouté aux données. Pour ce qui est de l'ajustement *vert* de la figure 2.4b, celui-ci semble être beaucoup plus représentatif des données que l'ajustement de la figure 2.4a. Les estimations des paramètres confirment cette observation, et il est intéressant de constater que l'EMQ est passée de 69.31 à 3.88 (Table III p.39).

Finalement, le modèle log-logistique à quatre paramètres est souvent perçu comme faisant abstraction au contexte expérimental du CHD. Je ne suis pas d'accord avec cela. Selon moi, cette approche est celle qui tient le plus compte du contexte expérimental tel qu'il ait. Les deux variantes décrites plus haut créent un contexte idéalistique et optimal bien que celui-ci ne soit pas observé dans les faits. Ajuster le modèle log-logistique à quatre paramètres permet d'observer le contexte expérimental plutôt que d'observer les effets d'un contexte imposé. Les deux ajustements de la figure 2.4b semblent bien représenter leurs données respectives. Les EMQs de ces ajustements sont aussi légèrement plus petite que celles pour la deuxième variante (Table III p.39). Les paramètres estimés sont aussi très semblables à ceux utilisés pour générer les données. Il est intéressant de noter que les paramètres a sont égales à -3.53 et -1.07 pour les ajustements *noir* et *vert* respectivement, bien que théoriquement ces valeurs devraient être de 0.00. Une inhibition négative de la croissance cellulaire suggère qu'il y aurait plutôt stimulation de la croissance. Dans le cas du paramètre a, cela indiquerait que pour une concentration infiniment petite d'un composé (voir en l'absence de composé), les cellules croissent. Pour des données synthétiques, une valeur négative est le résultat du bruit synthétique ajoutés au données ; pour des données expérimentales, une valeur négative est plutôt un artefact de la normalisation. Nous abordons d'avantage ce sujet dans la discussion du Chapitre 4.

Obtenir le maximum d'information d'un ajustement. Tel qu'attendu, le choix

(a) 1^e variante **(b)** 2^e variante **(c)** Modèle log-logistique

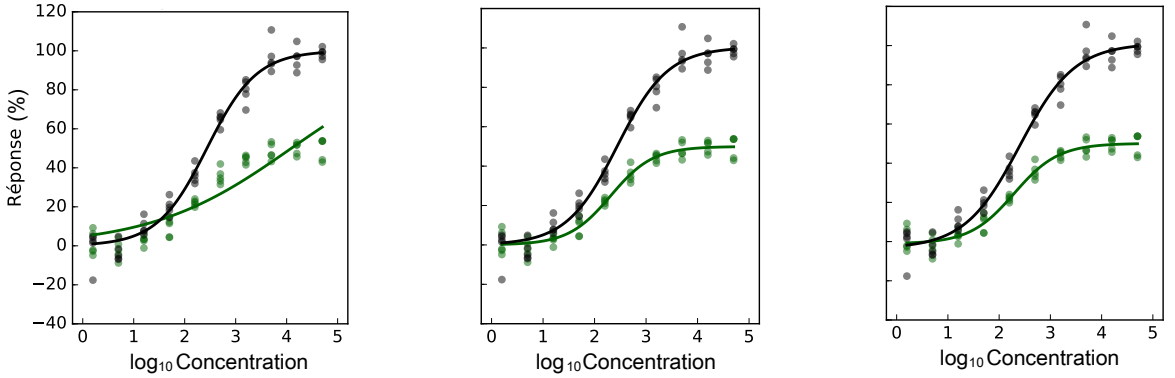


Figure 2.4 – Ajustements des variantes du modèle log-logistique. Les ajustements sont comparés pour deux jeux de données synthétiques légèrement bruitées ($\sigma = 5$). Les paramètres a, c et s sont sensiblement les mêmes pour les deux jeux. Ils diffèrent l'un de l'autre quant à la valeur du paramètre b : 100.00 pour le jeu *noir* et 50.00 pour le jeu *vert*. La table III identifie les estimations de paramètres pour les différents jeux et ajustements. 2.4a. Seuls les paramètres c et s sont ajustés. 2.4b. Les paramètres b, c et s sont ajustés. 2.4c. Tous les paramètres sont ajustés.

du nombre de paramètres à ajuster affecte les estimations des paramètres ainsi que la quantité et la qualité d'informations que l'on peut tirer d'un ajustement. Les conclusions tirées d'une analyse peuvent alors être largement erronées.

En fixant des paramètres à des valeurs constantes, l'ajustement est alors contraint à un plus petit nombre de dimensions diminuant ainsi la flexibilité de l'ajustement. Lorsque ces contraintes ne sont pas représentatives de la réalité des données analysées, les paramètres ajustés sont eux aussi non représentatifs des données.

La quantité d'information que l'on peut tirer d'un ajustement est propportionel au nombre de paramètres ajusté. Dans le cas de la première variante, nous étions seulement informés quant aux paramètres c et s et quant à l'EMQ. Les paramètres a et b étant constants, l'ajustement n'apportait aucune nouvelle information. Dans le cas du modèle log-logistique à quatre paramètres, nous avons alors de l'information pour tous les paramètres en plus de l'EMQ. De plus, l'ajustement du paramètre a peut être indicateur de problèmes expérimentaux. Par exemple, si $a = 25.00$ il pourrait y avoir eu contamination des contrôles négatifs ou une erreur lors du calcul de normalisation

L'information obtenue n'étant pas la même d'un ajustement à l'autre, les conclusions tirées seront elles aussi différentes. Notre jeu de données synthétiques *vert* est l'example parfait de cela. Cette variation dans les conclusions tirées est problématique car nous utilisons les paramètres ajustés comme métriques décrivant l'efficacité du composé analysé. Effectivement, les paramètres a et b représentent les réponses basale et maximale générées par le composé, le paramètre c représente l'IC_{50}, et le paramètre s est quant à lui indicateur de la rapidité des effets du composé. Utiliser le mauvais modèle confèrerait de fausses propriétés à un composé chimique.

Dans le but d'optimiser notre exploitation du modèle log-logistique lors de l'analyse de données de CHD, j'ai décidé d'utiliser uniquement le modèle à quatre paramètres.

Table III – Estimations des paramètres pour différents jeux de données et différentes variantes du modèle log-logistique

Paramètre	1^e **variante**		2^e **variante**		**log-logistique**	
	Noir	*Vert*	*Noir*	*Vert*	*Noir*	*Vert*
a	0.00	0.00	0.00	0.00	-3.53	-1.07
b	100.00	100.00	100.37	49.96	101.13	50.10
$\log_{10} c$	2.44	4.09	2.44	2.31	2.40	2.29
s	0.89	0.32	0.88	1.06	0.81	1.00
EMQ	34.07	69.31	5.84	3.88	5.70	3.85

2.4 Intervalles de confiance et fiabilité des estimations

Bien que l'ajustement du modèle log-logistique à quatre paramètres semble être plus représentatif de la réalité des données, les estimations des paramètres ne sont toujours que des *estimations*. Rappelons-nous que nous tentons de trouver un θ approximant $\theta_{\text{réel}}$. Le calcul d'un intervalle de confiance pour chacun des paramètres ajustés peut nous indiquer la fiabilité de nos estimations. Cet exercice est important puisque que les paramètres sont interprétés de telle sorte à caractériser l'efficacité d'un composé chimique.

Déterminer les limites de l'intervalle. Il est difficile de déterminer un intervalle de confiance pour un paramètre donné puisque nous n'avons que sa valeur. Nous devons dans un premier temps obtenir plusieurs valeurs pour ce paramètre. Pour ce faire, je suis retournée au contexte de la régression non-linéaire qui stipule qu'il existe un univers de jeux de données D_j pour $f(x; \theta_{\text{réel}})$. Les données utilisées pour faire l'ajustement ne représentent qu'un sous-ensemble des données associées à $\theta_{\text{réel}}$. Les estimations de paramètres que l'on obtiendrait en ajustant notre modèle aux différents D_j nous permettraient de calculer un intervalle de confiance pour les paramètres obtenus avec D.

Dans le but de générer des D_j représentatifs du contexte de l'analyse, j'ai exploré deux principales approches : la simulation *Monte-Carlo* (SMC) et le ré-échantillonnage *Bootstrap* (RB). Pour ces deux approches nous assumons que les paramètres estimés θ sont raisonnablement représentatif de $\theta_{\text{réel}}$ et que les erreurs aléatoires expérimentales

n'affectent pas cette supposition. Nous pouvons alors générer N nouvelles données depuis notre jeu initial D. Pour la SMC, un D_j est créé en substituant les valeurs de D par des valeurs aléatoirement sélectionées depuis une distribution donnée. La distribution choisie doit au meilleur de notre compréhension représenter le plus fidèlement possible la réponse étudiée [27]. Le nouveau jeu et les données obtenus par SMC sont tous deux synthétiques. Le RB est une alternative à la SMC lorsque nous détenons peu d'information concernant l'erreur à évaluer. Contrairement à la SMC, seul le nouveau jeu de données est synthétique : les données en tant que telles sont les mêmes que celles du jeu initial D. Un D_j est créé en sélectionnant aléatoirement et avec possibilité de répétions N données du jeu intial D [27].

Les D_j obtenus avec une ou l'autre des approches sont assujettis au même algorithme d'optimation que le jeu initial D. Ce processus est répété K fois et les estimations θ_k sont conservées. Il existe différentes techniques pour obtenir un intervalle de confiance à partir de ces données. Les deux approches les plus couramment utilisées sont celles des intervalles normaux et par centiles. L'approche des intervalles normaux nécessitent un large ensemble de θ_k et il est préférable que la distribution de ceux-ci soit normale. L'intervalle est symétrique autour de l'estimation initiale et se définit par le produit de la valeur normal $z_{\alpha/2}$ et de l'écart type des θ_k (Éq. 2.2). La moyenne des θ_k est dénotée par $\bar{\theta}$.

$$\theta_{100(1-\alpha)} = \theta \pm z_{\alpha/2} \sqrt{\sum_{k=1}^{K} \frac{(\theta_k - \bar{\theta})^2}{K-1}} \tag{2.2}$$

Les intervalles par centiles ne sont quant à eux pas nécessairement symétriques, puisque la distribution des θ_k n'a pas besoin d'être parfaitement normale. Les valeurs des θ_k sont ordonnées de façon croissante. Les limites de l'intervalle sont définies par les centiles $\alpha/2$ et $1-(\alpha/2)$ de l'ensemble croissant. La valeur de K doit être suffisamment large ($\leq 1\,000$) pour pouvoir générer des intervalles statistiquement fiables.

Puisque nous répétons K fois le processus d'ajustement, il est important que l'algorithme converge rapidement. Les gains en vitesse obtenus la section 2.2 sont alors très utiles.

Simulation et ré-échantillonnage. Pour déterminer laquelle des approches présentées plus haut est la mieux adaptée aux analyses faites dans le contexte du présent travail, j'ai étudié trois types de SMC ainsi que deux types de RB sur différents jeux de données synthétiques :

- *SMC spécifique à chaque concentration (SMC1).* La simulation se fait par concentration. Pour R réplicats, R données sont aléatoirement sélectionnées de la distribution $\mathcal{N}(f(x;\theta), \frac{\sum_{j=1}^{R}(y_j-\bar{y})^2}{R-1})$, où x est la concentration.

- *SMC avec erreur constante (SMC2).* La simulation se fait par concentration. Les données simulées proviennent de la distribution $\mathcal{N}(f(x,\theta), \sigma^2)$ pour une concentration x et un σ^2 constant.

- *SMC avec erreur constante selon les données (SMC3).* La simulation est similaire à la SMC2. La valeur de la variance est déterminée par $\frac{\sum_{j=1}^{C}\sum_{l=i}^{R}(y_l-\bar{y_j})^2}{N-1}$ où $\bar{y} = \frac{\sum_{j=1}^{R} y_j}{R}$, où C est le nombre de concentrations au total.

- *RB sur l'ensemble des réponses (RBR).* Le ré-échantillonnage se fait sur l'ensemble des réponses pour toutes concentrations confondues. Un cas extrême serait de générer un jeu de données ayant N fois la même observation.

- *RB sur les réplicats de chaque concentration (RBC).* Le ré-échantillonnage est fait par concentration. Un RBR est fait sur l'ensemble des réplicats d'une concentration.

J'ai testé ces approches sur différents jeux de données synthétiques. Ces jeux diffèrent sur deux points : (1) le nombre de réponses par concentration (réplicats) et (2) le bruit appliqué aux données ($\sigma = 0.2, 5, 10, 15$). Les résultats obtenus étaient très similaires : l'approche de simulation ou ré-échantillonnage doit être adaptées au nombre de réplicats

par concentrations. Je m'attarderai à l'analyse des résultats obtenus pour un jeu de données bruitées ($\sigma = 15$). Les approches de simulation et ré-échantillonnage furent répétés 1 000 fois avant de calculer les intervalles de croissance. Les jeux analysés ont aucun réplicat (une réponse par concentration), 2 ou 10 réplicats (Fig. 2.5 p.42).

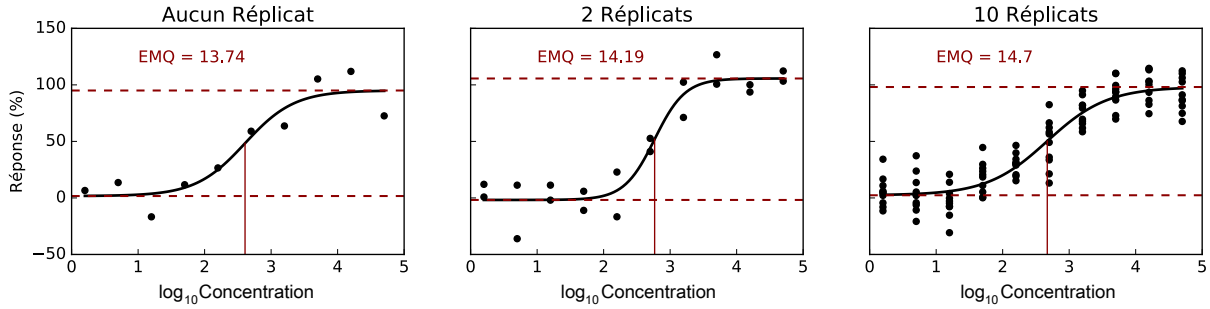


Figure 2.5 – Ajustements et EMQ selon le nombre de réplicats. Ajustements du modèle log-logistique à des données synthétiques bruitées ($\sigma = 15$) ayany aucun, 2 et 10 réplicats (points noirs). La valeur de l'EMQ est inscrite en rouge sur chacun des graphiques. Les tracés rouges hachés identifient les paramètres a et b des ajustements, tandis que le tracé rouge continu identifie le paramètre $\log_{10} c$. Les estimations de paramètres pour chacun des ajustements sont décrits dans la table IV

Table IV – Estimations des paramètres selon le nombre de réplicats

	a	b	$\log_{10} c$	s	EMQ
0 Réplicat	1.68	95.02	2.61	1.17	13.73
2 réplicats	-1.78	105.65	2.77	1.89	14.19
10 réplicats	2.36	98.11	2.67	0.96	14.70

Il n'est pas trivial d'évaluer et de comparer l'exactitude d'intervalles de confiance. J'ai tenté d'établir des intervalles *étalons* pour chaque jeu de données. Ces intervalles sont dérivés d'une SMC2 où la distribution utilisée pour simuler de nouvelles données et la même que celle utilisée pour générer les données synthétiques du jeu principale. Les intervalles obtenus devraient alors être rerprésentatifs de l'erreur sur les paramètres estimés lors de l'ajustement initial.

L'approche qui semble générer les intervalles les plus semblables aux étalons est la

simulation Monte-Carlo avec erreur constante (SMC2) lorsque $\sigma = EMQ$ (Fig. 2.6, 2.7 & 2.8 p.43-45). Il est intéressant de remarquer que les EMQs des ajustements initiaux sont très semblables et se rapprochent grandement du $\sigma = 15.00$ utilisé pour générer les jeux synthétiques (Table IV p.42). Cela n'est pas surprenant lorsque l'on compare l'équation de l'EMQ avec celle de l'écar-type : l'une utilise les résiduels tandis que l'autre utilise les écarts par rapport à la moyenne (Éq. 2.3). L'EMQ est généralement inférieure au bruit gaussien du jeu de données.

$$EMQ = \sqrt{\frac{\sum_{n=1}^{N}(y_n - f(x_n;\theta))^2}{N-1}} \approx \sqrt{\frac{\sum_{n=1}^{N}(y_n - \bar{y})^2}{N-1}} = \sigma \tag{2.3}$$

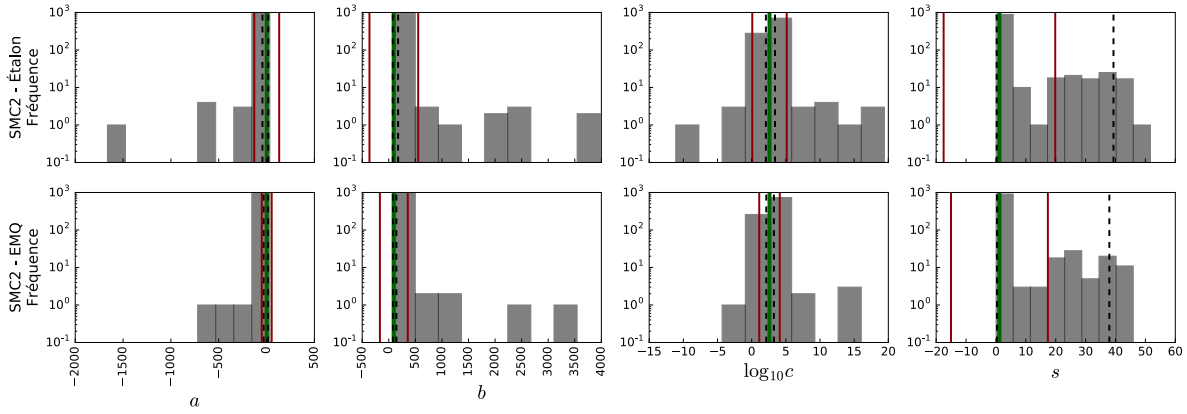


Figure 2.6 – Intervalles de confiance par simulation Monte-Carlo pour un jeu ayant aucun réplicat. Un histogramme est représentif de la distribution des valeurs d'un paramètre pour 1 000 SMC. Les estimations de l'ajustement initial sur un jeu de données synthétiques bruitées ($\sigma = 15$) sans réplicat sont marquées par des traits verts continus. Les limites des intervalles de confiance normaux sont marquées par les trais hachés rouges et celles des intervalles par centiles par des trais continus noirs. Les intervalles étalons se trouvent sur la première rangée. La deuxième rangée illustrent les résultats de l'approche par SMC2-EMQ.

Un avantage de cette approche est qu'elle ne semble par être affectée par le nombre de réplicats. Cela n'est pourtant pas le cas de la simulation Monte-Carlo avec erreur constante selon les données (SMC3), une variante de la SMC2. Le calcul de l'écart-type

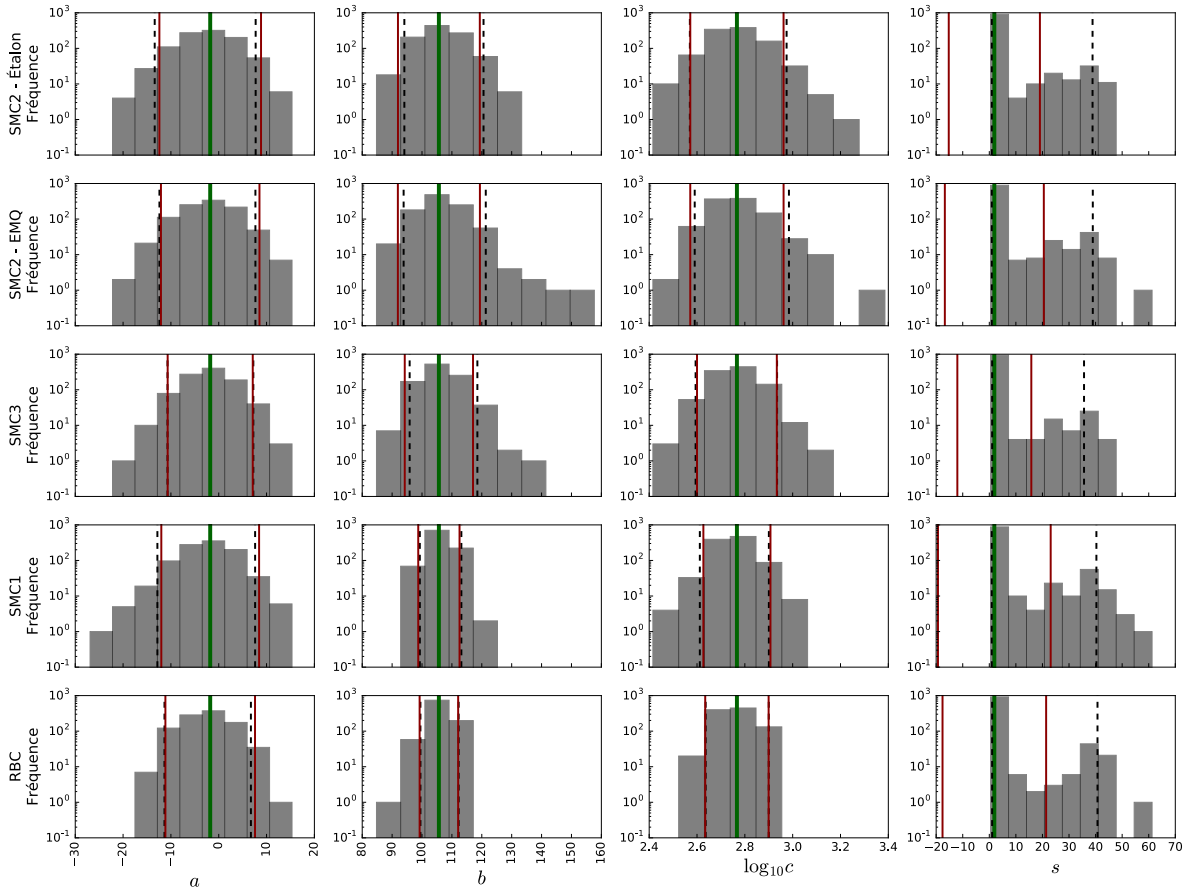


Figure 2.7 – Intervalles de confiance par simulation Monte-Carlo et ré-échantillonnage Bootstrap pour un jeu ayant deux réplicats. Un histogramme est représentif de la distribution des valeurs d'un paramètre pour 1 000 SMC ou RB. Les estimations de l'ajustement initial sur un jeu de données synthétiques bruitées ($\sigma = 15$) ayant 2 réplicats par concentration sont marquées par des traits verts continus. Les limites des intervalles de confiance normaux sont marquées par les trais hachés rouges et celles des intervalles par centiles par des trais continus noirs. Les intervalles étalons se trouvent sur la première rangée. Les rangées subséquentes illustrent les résultats pour les approches SMC2-EMQ, SMC3, SMC1, RBC.

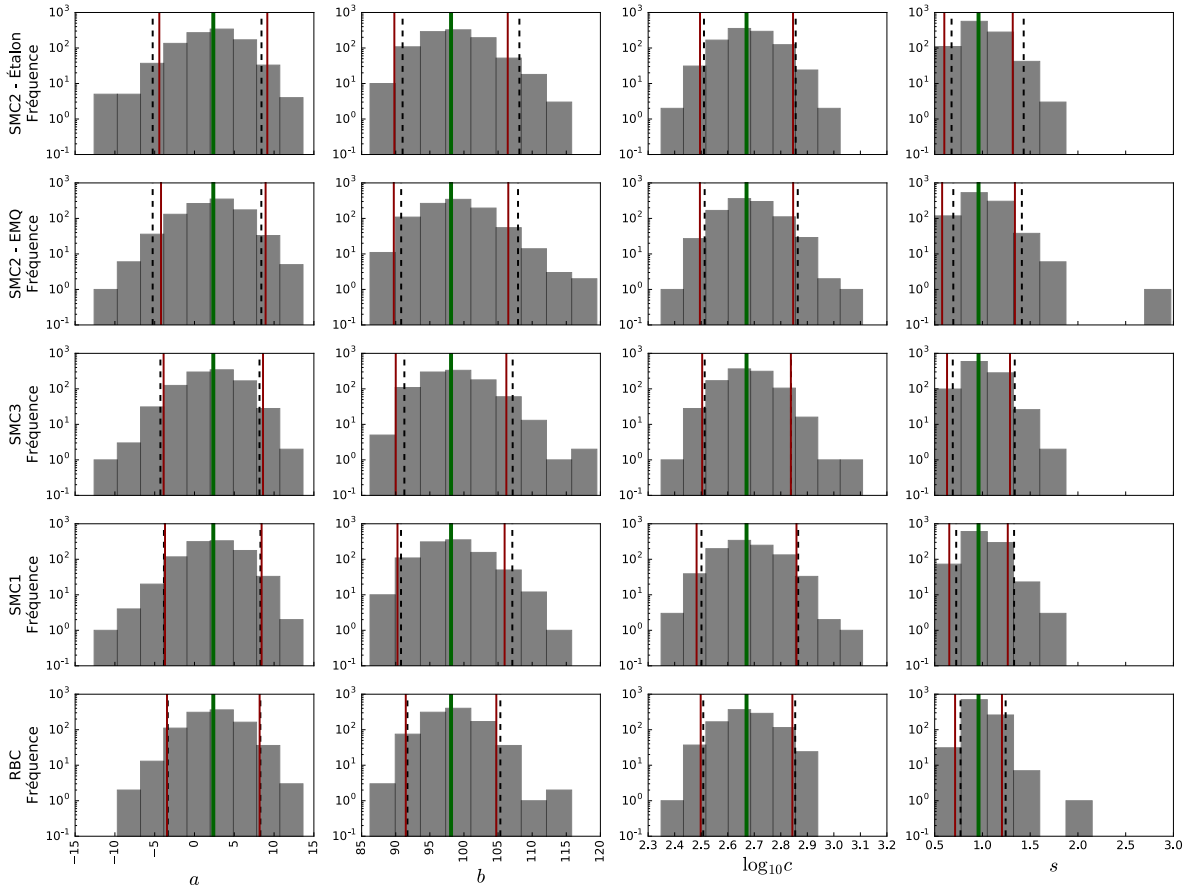


Figure 2.8 – Intervalles de confiance par simulation Monte-Carlo et ré-échantillonnage Bootstrap pour un jeu ayant deux réplicats. Un histogramme est représentif de la distribution des valeurs d'un paramètre pour 1 000 SMC ou RB. Les estimations de l'ajustement initial sur un jeu de données synthétiques bruitées ($\sigma = 15$) ayant 10 réplicats par concentration sont marquées par des traits verts continus. Les limites des intervalles de confiance normaux sont marquées par les trais hachés rouges et celles des intervalles par centiles par des trais continus noirs. Les intervalles étalons se trouvent sur la première rangée. Les rangées subséquentes illustrent les résultats pour les approches SMC2-EMQ, SMC3, SMC1, RBC

de la distribution normale utilisée lors de la simulation est grandement affecté par le nombre de réplicat. En l'absence de ceux-ci, le calcul est tout simplement inutile puisque $y_j = \bar{y}_j$. En présence de deux réplicats, bien que l'écart-type soit calculable, il n'est pas nécessairement représentatif de l'erreur. Le nombre de réponses totales ne semblent pas être suffisant pour estimer un écart-type représentatif du bruit initial (Fig. 2.7 p. 44). Par exemple, l'écart-type calculé pour le jeu de données à deux réplicats de la figure 2.5 est de 12.88. Si l'on tente de généraliser et que l'on calcule la racine de la variance moyenne pour 1 000 jeux de données synthétiques bruitées ($\sigma = 15$) ayant deux réplicats par concentration, on remarque que sa valeur de 10.02 ne représente pas le bruit initial. Cela étant dit, le calcul de l'écart-type augmente de précision lorsqu'il y a plus de réplicats par concentration. La racine de la variance moyenne pour 1 000 jeu de données synthétiques bruitées ($\sigma = 15$) ayant maintenant dix réplicats par concentrations est de 14.24. Cette valeur se rapproche d'avantage du bruit connu $\sigma = 15.0$. Les intervalles obtenus sont alors semblables à ceux obtenus avec la SMC2 où $\sigma = EMQ$ (Fig. 2.8 p. 45). Or, l'utilisation de dix réplicats ou plus n'est pas la norme lors des expériences de CHD : cela est très coûteux en terme d'échantillons. Normalement, les expériences contiennent deux réplicats par concentration, et même parfois aucun. Dans ces deux cas, la SMC3 ne semble pas être une approche optimale. De plus, il reste à savoir si l'erreur est le même pour l'ensemble des données. Dans le cas des données synthétiques présentées dans la figure 2.5. nous savons que c'est le cas. Cependant, cela n'est pas certain lorsque nous analysons des données expérimentales. Je discuterai d'avantage de ce point dans le Chapitre 3.

Une alternative à la SMC3 pour lorsque le bruit n'est pas le même sur l'ensemble des données est la simulation Monte-Carlo spécifique à chaque concentration (SMC1). Cette approche n'est toute fois pas applicable lorsqu'il n'y a aucun réplicat : il est mathématique impossible de calculer l'écart-type de $\mathcal{N}$ lorsque $R - 1 = 0$. Dans le cas où il y a deux réplicats, le calcul se fait mais manque encore de précision : deux valeurs ne sont pas suffisantes pour déterminer adéquatement un écart-type. Il y a beaucoup de variation dans les valeurs calculées pour différentes concentrations. Pour le jeu de données synthétiques à deux réplicats présenté dans la figure 2.5, seulement deux concentrations ont des valeurs

générant un écart-type semblables au bruit initial, soit les concentrations 3.2 et 3.7 avec des écarts-types de 15.59 et 13.01 respectivement (Fig. 2.9 p.47). La variation dans les écarts-types est moins importantes lorsqu'il y a dix réplicats (Fig. 2.9 p.47). Les valeurs se rapproches aussi d'avantage du bruit initial. Comme pour la SMC3, la SMC1 semble peut efficace lorsqu'il n'y a que très peu de réplicats, ce qui est généralement le cas des données expérimentales. Les intervalles obtenus ressemblent très peu aux étalons et à ceux obtenus par SMC2 lorsque $\sigma = EMQ$ (Fig. 2.7 p. 44). Cependant, en présence de plusieurs réplicats (≥ 10), les intervalles sont semblabes aux étalons : les approches SMC2 avec $\sigma = EMQ$, SMC3 et SMC1 génèrent des intervalles très semblables (Fig. 2.8 p. 45).

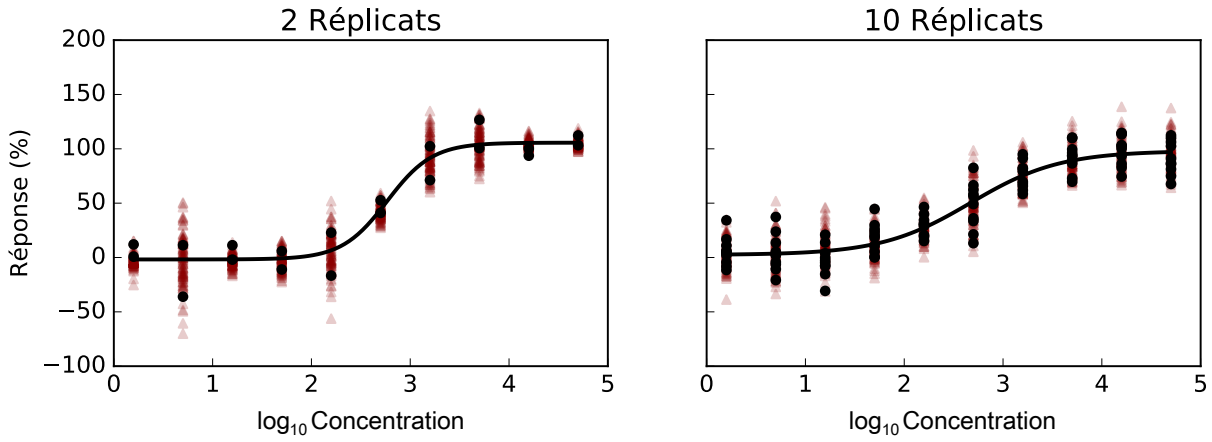


Figure 2.9 – Représentation des écart-types de la SMC1 selon le nombre de réplicats. La SMC1 calcul un écart-type par concentration x selon les données synthétiques représentées par les points noirs. L'écart-type est utilisé pour générer une distribution normale ayant comme valeur centrale $f(x;\theta)$. Les triangles rouges représentent 250 valeurs aléatoires de cette distribution.

Les approches de simulation Monte-Carlo semblent bien fonctionner lorsqu'il y a plusieurs réplicats dans notre jeu de données. Les intervalles générés par ces approches semblent bien estimer la fiabilité des paramètres ajustés. Ces approches semblent cependant moins bien performer lorsqu'il y a peu ou absence de réplicat, faute d'information disponible pour caulculer l'écart-type de la distribution modélisant l'erreur. Bien que plus de réplicats aide à mieux déterminer cette distribution, reste encore à savoir si l'erreur est

commune et si elle est normale. Dans le cas des jeux de données synthétiques utilisés dans cette section, le bruit était commun et normal. Cela explique selon moi la similarité entre les intervalles pour les différentes approches lorsqu'il y a dix réplicats. Dans le cas de données expérimentales où nous ne connaissons pas le bruit *(a priori*, les résultats pourraient être différents. Je discuterai d'avantage de cette problématique dans le Chapitre 3.

Le ré-échantillonnage Bootstrap est une bonne alternative aux simulations Monte-Carlo lorsque les données ne sont pas assez informatives ou lorsque le bruit sur les données n'est pas normal. L'approche par Bootstrap, comparé aux SMCs, ne dépendant d'aucune distribution : elle évalue indirectement la distribution des données. Dans le cadre du présent travail, j'ai testé deux approches de ré-échantillonnage. La première, le ré-échantillonnage Bootstrap sur l'ensemble des réponses (RBR) est inadéquate. Les jeux de données obtenus ne sont pas représentatifs du contexte étudié. Par exemple, certains jeux pourraient avoir moins de concentrations que le jeu initial. Il se pourrait aussi qu'il y ait absence de réponses minimales et/ou maximales. Les paramètres a et b seraient alors ajustés à des valeurs nos représentatives de l'expérience analysée. Les jeux de données obtenus par ré-échantillonnage devrait avoir au moins une réponse par concentration testée. J'ai implémenté un ré-chantillonage Bootstrap par concentration (RBC). Cette approche ressemble à la SMC1 et est donc aussi inaplicable lorsqu'il y a absence de réplicat. Pour peu de réplicat, les intervalles obtenus sont très semblables à ceux obtenus avec la SMC1 (Fig. 2.7 p. 44). Pour ce qui est des expériences avec un grand nombre de réplicats, les intervalles sont un tant soit peu optmistiques comparés aux étalons, mais leurs valeurs ne divergent pas trop de celles des SMCs (Fig. 2.8 p. 45).

Généralisation du calcul des intervalles. Le calcul d'intervalles de confiance pour les différents paramètres estimés lors de la régression non-linéaire semble bien se faire par simulation Monte-Carlo avec erreur constante en utilisant la distribution $\mathcal{N}(f(x;\theta), EMQ)$. Cette approche est inaffectée par le nombre de réplicats. Gardons en tête que les jeux de données synthétiques utilisés avait tous un bruit normal commun, ce qui n'est pas toujours le cas lorsque l'on analyse des données expérimentales. Les approches par simulation

Monte-Carlo nécessitent une certaine compréhension du bruit sur les données et pourraient ne pas convenir lors d'analyses expérimentales. L'approche par ré-échantillonnage Bootstrap semble alors une bonne alternative. Nous devons cependant s'assurer que les jeux de données obtenus respectent le contexte expérimental, c'est-à-dire qu'il y ait le même nombre de concentrations que dans le jeu initial. Je propose d'utiliser un ré-échantillonnage Bootstrap par concentration. Dans le prochain chapitre, nous testerons et discuterons de ces approches lorsqu'elles sont appliquées à l'analyse de données expérimentales.

Les intervalles obtenus sont de bon indicateurs de la stabilité des paramètres estimés. Nous remarquons que les distributions des différentes valeurs obtenues par simulation ou ré-échantillonnage ne sont pas toujours normales (Fig. 2.6, 2.7 & 2.8 p.43-45). Le calcul des intervalles par centiles semble alors le plus approprié. Dans les cas où la distribution est gaussienne, les intervalles normaux et par centiles sont très similaires.

Rappelons-nous que le but du présent travail est de mettre sur pied un processus d'analyse automatisé. Celui-ci doit donc être applicable pour un très grand nombre de contextes. Les analyses décrites plus haut ont permis de cerner les avantages et inconvénients des approches testées. D'autres analyses devront cependant être faites avant de déterminer une approche unique, optimisant le calcul d'intervalles de confiance pour la majorité des contextes d'analyse.

2.5 Interprétation et comparaison des ajustements

Différentes analyses et interprétations peuvent être faites depuis les résultats d'un ajustement. Tel que mentionné plus ahut, il est possible de caractériser un composé en interprétant les paramètres estimés. Il est aussi possible de dériver d'autres métriques, telle que l'aire sous la courbe dose-réponse (AUC[4], de l'anglais "*area under the curve*") [41, 42, 43]. L'AUC se calcule en intégrant notre modèle log-logistique ajusté entre deux valeurs de x

[4] À ne pas confondre avec l'AUC d'une courbe ROC

(Éq. 2.4).

$$AUC = \int_{x_1}^{x_2} f(x;\theta)dx \tag{2.4}$$

Les unités de l'AUC sont égales au produit les unités des concentrations et des réponses. Cela étant dit, il est très difficile d'interpréter ces unités. Par exemple, l'ajustement A de la figure 2.10 a une AUC de 118.34% par unité de concentration. Lorsque l'on compare les AUCs de deux ajustements, il est important qu'elles aient été calculées pour les mêmes valeurs de x_1 et x_2. Ces valeurs sont généralement les plus petite et plus grande concentration utilisées lors de l'expérience.

L'attrait de l'AUC réside dans le fait que cette seule valeur combine les effets des paramètres b et c : plus le paramètre b est large, plus l'AUC sera grande (Fig. 2.10 p.51) ; plus le le paramètre c est petit, plus l'AUC sera grande (Fig. 2.11 p.52).

Table V – Estimations des paramètres utilisés lors du calcul des AUCs pour les ajustements A, B, C et D

	a	b	$\log_{10} c$	s
Ajustement A	-1.07	**50.10**	2.29	1.00
Ajustement B	-3.53	**101.12**	2.40	0.80
Ajustement C	2.61	100.00	**1.75**	0.99
Ajustement D	0.68	98.50	**2.88**	1.10

Bien que l'AUC soit couramment utilisée, son calcul reste parfois problématique. Que faire lorsque le paramètre a est négatif et que la réponse à x_1 soit elle aussi négative ? Nous devons alors modifier notre x_1. Cependant, lorsque l'on compare deux ajustements, cette nouvelle valeur doit être adéquate pour le calcul des deux AUCs. Une problématique similaire survient lorsque la valeur du paramètre a n'est pas atteint dans la gamme des concentrations expérimentales (Fig. 2.11 p.52). Nous discuterons d'avantage de l'AUC dans le Chapitre 4.

Une autre façon d'analyser et d'interpréter les données de CHD est en comparant deux

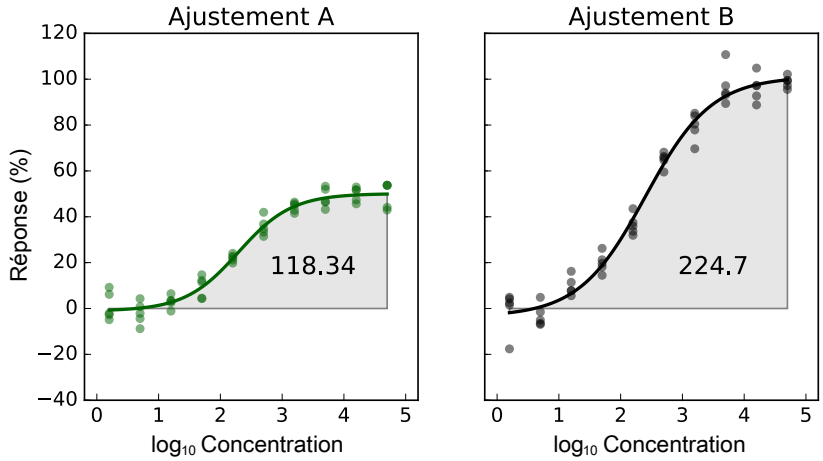


Figure 2.10 – Effets du paramètre b sur l'AUC. Les jeux de données synthétiques légèrement bruitées ($\sigma = 5$) sont obtenus pour les mêmes paramètres $a = 0.00$, $\log_{10} c = 2.50$ et $s = 1.00$. Les données de l'ajustement A sont obtenues pour $b = 50.00$ tandis que celle de l'ajustement B sont obtenues pour $b = 100.00$. Les paramètres estimés des deux ajustements sont reportés dans la table V. L'AUC est représentée par la partie ombragée sous les courbes, et sa valeur est indentiquée sur chaque graphique.

.

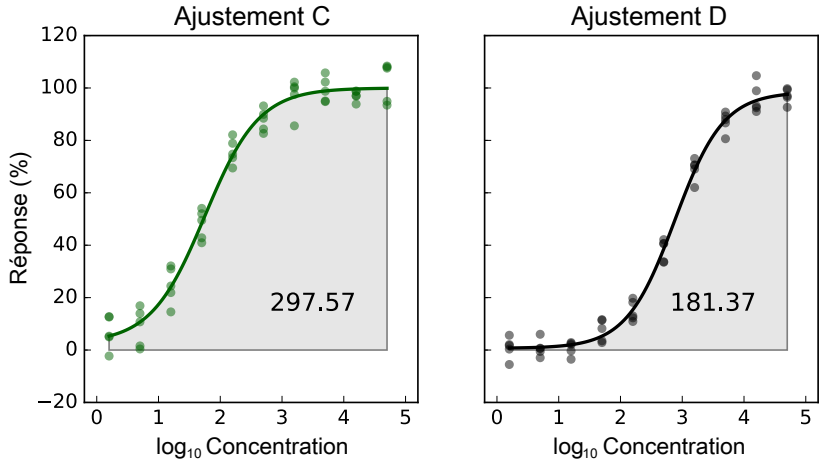


FIGURE 2.11 – EFFETS DU PARAMÈTRE c SUR L'AUC. Les jeux de données synthétiques légèrement bruitées ($\sigma = 5$) sont obtenus pour les mêmes paramètres $a = 0.00$, $b = 100.00$ et $s = 1.00$. Les données de l'ajustement A sont obtenues pour $\log_{10} c = 1.70$ tandis que celle de l'ajustement B sont obtenues pour $\log_{10} c = 2.90$. Les paramètres estimés des deux ajustements sont reportés dans la table V. L'AUC est représentée par la partie ombragée sous les courbes, et sa valeur est indentiquée sur chaque graphique.

ajustements. Dans le contexte du développement de médicament, nous sommes souvent intéressés à comparer les effets d'un composé sur différents types cellulaires ou sur différents patients (voir la Chapitre 3 pour plus de détails sur la notion de *patient*), ainsi qu'à comparer les effets de deux composés sur un même type cellulaire ou sur un même patient. Plusieur comparaisons se font "à l'oeil" sans valeur statistique. Une autre approche couramment utilisée est de tout simplement comparer les valeurs estimées des IC_{50} : le composé ayant le plus petit des IC_{50} est libellé comme étant le plus "*efficace*" des deux. Je propose ici une approche mathématique pour comparer deux ajustements et ce, en considérant tous les paramètres estimés lors de l'ajustement.

Pour un paramètre donné p, nous pouvons établir une relation de grandeur entre les estimations p_A et p_B où A et B sont différents ajustements. La fréquence à laquelle nous observons l'inverse de cette relation dans les données de simulation ou ré-échantillonnage représente la p-value de notre test statistique. Je dis d'une valeur de p qu'elle est signifi-

cativement supérieure à une autre lorque notre p-value est égale ou inférieure à 0.01. Cela signifie que pour les K estimations de p obtenues par simulation ou ré-échantillonnage, la valeur évaluée est inférieur à l'autre dans seulement 1% des cas.

Prenons les ajustements A et B de la figure 2.10. Les données synthétiques utilisées pour ces ajustements sont générées depuis deux modèle très similaires. Ceux-ci partagent les paramètres $a = 0.00$, $\log_{10} c = 2.70$ et $s = 1.00$. Le paramètre b est de 50.00 pour le jeu de l'ajustement A et de 100.00 pour le jeu de l'ajustement B. À ces modèles, un bruit gaussien ($\sigma = 5$) est ajouté pour générer les données telles que présentées dans la figure 2.10. Une fois l'ajustement initial fait, la SMC2 avec $\sigma = EMQ$ est appliquée aux deux ajustements pour 1 000 répétitions. Je compare ensuite les ajustements pour tous les paramètres θ. Tel qu'attendu, les ajustements A et B diffèrent de par leur paramètre b : la réponse maximale de l'ajustement B est significatement plus élevée que celle de l'ajustement A. Il est intéressant de constater qu'il semble avoir une certaine distinction entre les valeurs des paramètres $\log_{10} c$ et s. Les nuages de points et les distributions des valeurs illustrent clairement les différences (et ressemblances) entre les valeurs des paramètres des deux ajustements semblent tout de mêmes se distinguer un tant soit peu (Fig. 2.12 p.54).

Cette expérience contrôlée démontre bien l'importance de considérer d'autres paramètres que l'IC_{50} lors de l'analyse de données de CHD. Nous aborderons davantage ce sujet dans le Chapitre 3 lors de l'analyse de données expérimentales.

2.6 Analyse dite de *groupe*

Le dernier aspect étudié est l'analyse dite de groupe. Les données analysées dans ce travail sont représentatives des effets d'un composé donné sur les cellules d'un patient (voir le Chapitre 3 pour plus de détails). Un ajustement est donc le portrait d'une relation composé-patient. Plusieurs expériences de criblage sont conçues de telle sorte à analyser

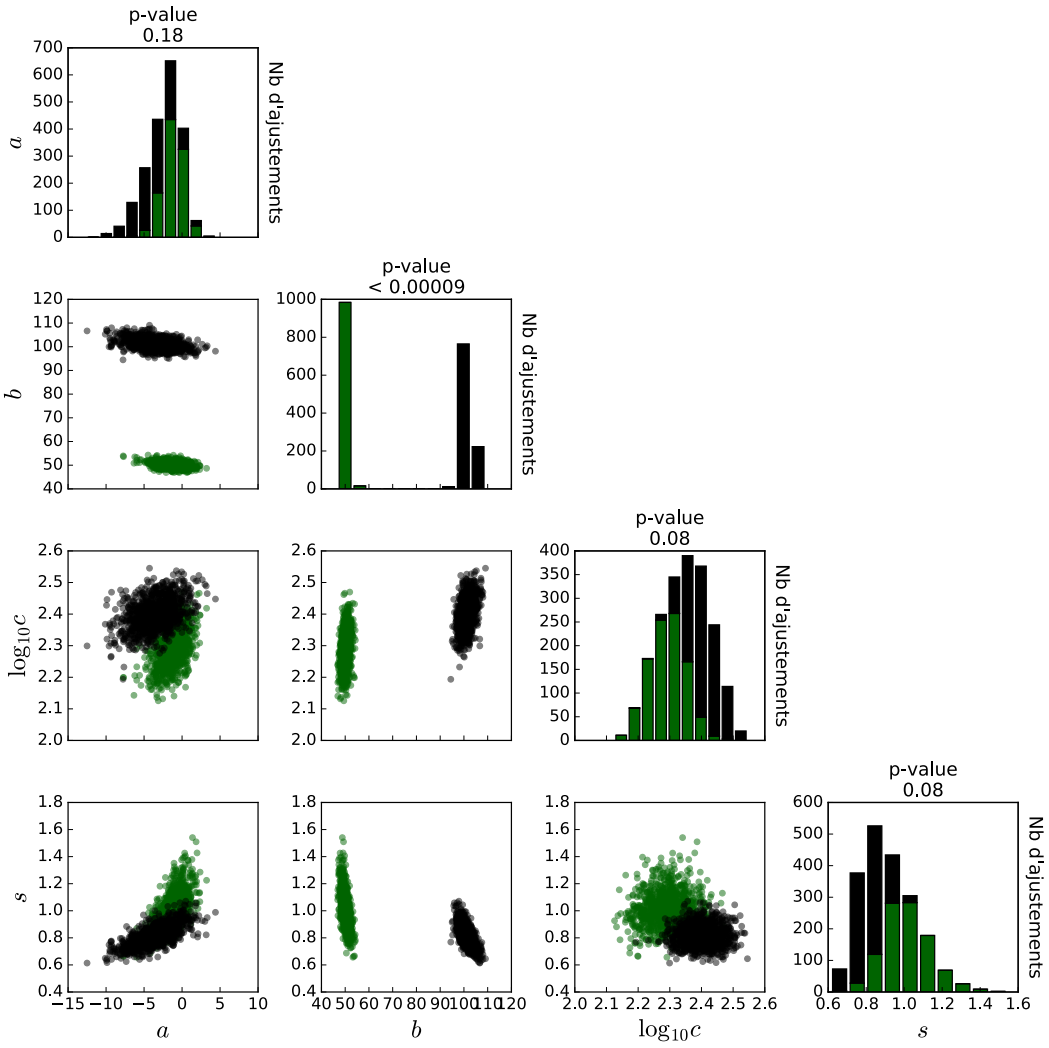


Figure 2.12 – Comparaison des ajustements A et B selon les données de SMC2. Les histogrammes illustrent les distributions des valeurs de chaque paramètre pour 1 000 SMC2. Les nuages de points illustrent la relation entre deux paramètres. Une p-value est calculée pour chaque paramètre et est inscrite au-dessus des histogrammes. Celle-ci détermine si la valeur d'un ajustment est significativement supérieure celle d'un autre ajustement. Les valeurs de l'ajustement A sont représentées en vert tandis que celles de l'ajustement B sont en noir. Ces ajustements diffèrent clairement au niveau du paramètre b.

différents groupes (ou parfois différents sous-groupes) de patients dans le but d'identifier des composés qui pourraient être spécifiques à un ensemble de patients. Dans cette optique, j'ai voulu élargir le processus d'ajustement à l'analyse de ces ensembles.

Un *groupe* se définit par un regroupement de P patients ayant tous été testés pour au moins un composé commun. Plutôt que d'analyser les ajustements individuels et tenter identifier des ressemblances, je propose de faire qu'un seul ajustement qui sera alors représentif d'un groupe pour un composé donné. Chaque patient est dès lors considéré comme un cas extrême de réplicats biologiques. Pour P patients ayant chacun R réplicats par concentration, le nombre d'observations utilisées pour l'ajustement devient $P \times R \times C$ (Fig. 2.13 p.56). Il est intéressant de noter que pour l'exemple de la figure 2.13, l'ajustement de groupe est équivalant à l'ajustement des moyennes (Table VI p.55). L'unique différence entre ces ajustements réside dans les valeurs de l'EMQ, ce qui n'est pas surprenant compte tenu du nombre d'observations utilisées pour chaque ajustement. Notons ici que les deux approches tiennent compte de la variance des réponses, mais de différentes façons ce qui explique aussi les différences dans les valeurs des EMQ. Cette équivalence des approches n'est cependant pas acquise. Les ajustements pour un groupe dont les patients sont testés pour différentes concentrations ne seront pas les mêmes. La différence dans les estimations sera aussi plus importante si le nombre de réponses par concentration n'est pas toujours le même.

Table VI – Estimations des paramètres pour un ajustement des moyennes et un ajustement de groupe

	a	b	$\log_{10} c$	s	EMQ
Ajustement Moyen	-0.16	90.55	2.88	0.89	1.22
Ajustement de Groupe	-0.16	90.55	2.88	0.89	19.93

L'ajustement des moyennes nous limite quant à la continuité de notre analayse. Nous avons observé plus haut qu'un plus grand nombre de réplicats aide à mieux établir les intervalles de confiance sur les paramètres estimés. Or, l'ajustement des moyennes se fait sur un jeu n'ayant qu'une réponse (moyenne) par concentration. Il sera alors difficile d'évaluer la fiabilité des paramètres estimés. En faisant plutôt un ajustement de groupe, nous pou-

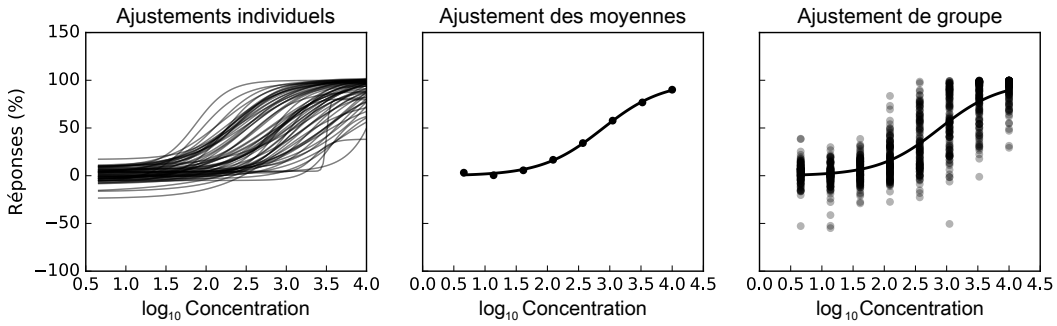


Figure 2.13 – Ajustements pour l'analyse d'un groupe. L'approche par ajustements individuels agrège les patients en groupe une fois la paramètrisation faite. L'approche par ajustement des moyennes regroupe les patients avant l'ajustement. La paramétrisation se fait sur les réponses moyennes de chaque concentration. L'ajustement de groupe se fait sur l'ensemble des réponses d'un groupe de patients.

vons calculer plus faiclement ces intervalles de confiance. De plus, ayant plusieurs réplicats par concentration, nous pouvons faire certains tests pour mieux déterminer laquelles des approches de simultaions ou de ré-échantillonnage conviendrait le plus.

En l'absence de normalité, les simulations Monte-Carlo ne sont pas entièrement représentatives de l'expérience analysées et les intervalles de confiance risques d'être incorrectes. Cependant, le ré-échantillonnage Bootstrap par concentration semble dans le contexte de groupe ne pas être une alternative. Cette approche "brise" le concept du regroupement de patients. Dans un cas extrême, un ré-échantillonnage pourrait générer un jeu où toutes les réponses de la première concentration proviennent du patient P^*. Dans un le contexte expérimental de l'analyse, un tel jeu de données n'est pas réaliste. En plus de resoecter le nombre de concentrations initial (voir section 2.4), le ré-échantillonnage doit aussi respecter notre concept du groupe de patients. Je propose alors d'utiliser un ré-échantillonnage sur les patients. Lorsqu'un patient est échantillonné, toutes ses réponses sont prises en considérations : inversement, lorsqu'il n'est pas échantilloné, aucune de ses réponses n'est considérée. Cette méthode de ré-échantillonnage nous permet d'évaluer la fiabilité des paramètres estimés par rapport au groupe.

Un autre aspect intéressant quant à l'analyse de groupe est la comparaison de deux groupes. En utilisant l'approche statistique décrite à la section 2.5, nous pouvons comparer un même groupe de patients pour différents composés ou deux groupes de patients pour un même composé. Le Chapitre 3 traitera de ce sujet dans un contexte expérimental bien précis.

2.7 Conclusion

Dans le but de mettre sur pied une approche automatisée, flexible et statistique pour l'analyse des données de CHD j'ai cerné puis étudié différents aspects du processus. Premièrement, l'algorithme Levenberg-Marquardt converge rapidement de façon générale bien qu'il est possible d'obtenir un léger gain en initiant adéquatement les méta-paramètres. Il semble aussi plus efficace d'initier les paramètres du modèle log-logistique à des valeurs dérivées des données évaluées lors de l'ajustement, plutôt que de façon arbitraire. Deuxièmement, il vaut mieux ajuster les quatre paramètres du modèle sans faire de suppositions *a priori*. Troisièment, il est possible de calculer des intervalles de confiance pour les paramètres estimés en faisant plusieurs simulations de données ou un ré-échantillonnages. L'approche choisie dépend du nombre de données et de nos connaissances quant au bruit sur les données. Quatrièmement, ces méthodes permettent aussi d'établir statistiquement si la valeur d'un paramètre pour un ajustement est significativement supérieure à la valeur du même paramètre mais pour un autre ajustement. Finalement, le processus peut être élargi à l'analyse de groupe.

Les différentes composantes du processus décrites dans ce chapitre seront appliquées à une étude expérimentale dans le Chapitre 3, soit l'étude des effets de différents composés sur des cellules leucémiques.

Chapitre 3

Étude comparative des effets de composés chimiques sur des cellules leucémiques

Le Chapitre 2 présentait les différents aspects considérés et étudiés lors de la mise en place du processus d'analyse. Le présent chapitre vise à démontrer l'application de ce processus dans un cadre expérimental. Ce travail fut possible grâce à une collaboration avec le laboratoire du Dr. Guy Sauveau de l'Institut de Recherche en Immunologie et Cancérologie (IRIC, Montréal). Les travaux et résultats présentés dans ce chapitre furent développés et obtenus dans le cadre du projet Leucégène. Ce projet interdisciplinaire vise à l'amélioration des soins de la leucémie myéloïde aiguë en préconisant une thérapie ciblée aux patients (médecine personnalisée).

3.1 Leucémie myéloïde aiguë

La leucémie myéloïde aiguë (LMA) est un groupe hétérogène de plusieurs leucémies. Différentes lignées cellulaires précurseurs seraient à son origine : myéloïde, érythroïde, mégacaryocytes et monocyraires [44]. Elle est causée par l'incapacité des ces précurseurs hématopoïétiques à se différencier en cellules matures et fonctionnelles. Dès lors, il y a accumulation de cellules immatures inhibées à différents stades de la différentiation, diminuant ainsi grandement la production normale d'éléments hématopoïétiques [45]. Les cellules leucémiques souches (CLS) détiennent des certaines anormalités génétiques et chromosomiques qui nuisent à leurs mécanismes de différentiation et d'apoptose [44], et d'autres qui avantagent leurs capacités de prolifération et d'auto-renouvellement [46]. Chez un patient, la LMA est maintenue par ces CLS, les blastes leucémiques ne se proliférant que très peu.

La LMA est généralement diagnostiquée ches des patients de 60 ans et plus, bien qu'elle peut être présente chez de jeunes patients [47, 44]. Bien que l'âge soit un facteur important lors du diagnostic et de la sélection du traitement, les profils cytogénétique et épigénétique du patient sont aussi à considérer. Ceux sont même souvent informateurs quant au pronostic. Des études ont démontré l'implication de certains motifs de cytosines méthylées lors du diagnostic de la LMA [48, 49]. Un traitement dit *déméthylant* serait alors approprié pour les patients présentant de telles signatures. D'autres études ont quant à elles démontré une corrélation entre l'âge du patient et certaines caractéristiques des profils mentionnés plus. Notamment, la LMA chez les personnes âgées ($>$ 60 ans) serait multi-résistante à différents médicaments et ces patients répondraient donc moins bien aux traitements de chimiothérapie [50, 51, 52]. Un autre exemple de cette relation âge-profil sont les mutations au gène KIT, un récepteur de tyrosine kinase. Ces mutations ne semblent être présentent que chez les adultes et sont associées à un haut risque de *rechute* : 70% des patients ayant le gène KIT muté développent une seconde LMA après une première rémission. Les adultes seraient alors plus susceptibles aux *rechutes* ainsi qu'à

un pronostic défavorable [50, 47].

La LMA est généralement traitée par chimiothérapie. La formule dite standard est sept jours de cytarabine et trois jours de daunorubicine. Chez les patients de moins de 60 ans, entre 70 et 80% atteindront une rémission complète (RC) avec ce traitement. Cependat, seulement 40 à 50% des patients de 60 ans et plus atteindront une RC pour ce même traitement. De ceux-ci, environ 85% développeront une seconde LMA dans les deux à trois anns suivant leur rémission. Cette nouvelle LMA est souvent encore plus résistante que la première et les pronostics ne sont que très peu favorables. De tels statistiques démontrent bien que les traitements actuels sont inaptes à iradiquer complètement la maladie chez le patient. Bien que très peu de nouveaux médicaments ont récemment été approuvés dans le contexte de la LMA [47], de présentes études visent à développer de nouveaux traitements plus efficaces [53, 8]. Ces études prennent de plus en plus la voie de la médecine personalisée, c'est-à-dire l'identification d'une thérapie effective pour un sous-groupe de patients.

3.2 Données expérimentales

Les données analysées proviennent toutes d'une même expérience de criblage. Une vingtaine de composés chimiques ont été testés sur des cellules leucémiques primaires de plus de 200 patients. Chaque composé était testé pour huit concentrations allant de 4.57 nM à 10 000.00 nM (4.57, 13.72, 41.15, 123.45, 370.35, 1 111.09, 3 333.3, et 10 000.0 nM). Pour des questions de confidentialité et puisque le but premier du travail est de développé un processus d'analyse, l'identification des composés et patients sera anonymisée.

Le CellTiter® Glow Assay (Promega) fut utilisé pour quantifier les réponses cellulaires aux différents composés et concentrations. Pour chaque ensemble patient-composé-concentration, une valeur numérique de luminescence est obtenue. Ces valeurs sont enregistrées dans une base de données (MongoDB) conçue par la plateforme de Bio-Informatique

de l'IRIC. Les données de luminscence (l_i) sont normalisées par la moyenne des contrôles négatifs ($\overline{\text{ctrl\_}}$) (Éq. 3.1). Les valeurs normalisées sont dès lors représentatives du taux (%) d'inhibition de la croissance cellulaire (y_i). Ce sont ces valeurs qui sont utilisées par le processus d'analyse et l'efficacité d'un composé est donc évaluée selon ce taux.

$$y_i = 100 - 100 \times \frac{l_i}{\overline{\text{ctrl\_}}} \tag{3.1}$$

Deux composés (*Composé1* et *Composé2*) d'intérêt sont analysés dans deux contextes différents : (1) l'analyse de patients individuels (*PatientA* et *PatientB*), et (2) l'analyse de groupes de patients (*GroupeA* et *GroupeB*). Ces composés ont été sélectionnés par mes collaborateurs selon les résultats d'expériences préliminaires.

3.3 Utilisation du processus d'analyse

Tel que mentionné plus haut, les données expérimentales utilisées se trouvent dans une base de données. Le processus d'analyse proposé au Chapitre 2 a été construit de telle sorte à se connecter sur une telle base de données (j'utilise l'outil PyMongo pour faire cela). Aucune manipulation de données externe au processus n'est fait : la normalisation, l'ajustement et les différents analyses sont tous faits de façon automatisée.

Le processus prend en entrée une liste de patients, de composés et d'expérience. Le formatage des données est fait selon le type d'analyse choisie, c'est-à-dire *par patient* ou *par groupe*. La méthode pour le calcul des intervalles doit aussi être spécifiée. Je travaille cependant à incorporer la sélection de cette méthode à même le processus d'analyse.

Tous les résultats présentés dans les prochaines sections ont été obtenus en utilisant le *processus d'ajustement par défaut* (PAD) tel que défini dans le Chapitre 2 : les méta-

paramètres de la regression non-linéaire sont $\mu_0 = 1.0$, $\lambda_1 = 10$ et $\lambda_2 = 100$; le modèle log-logistique ajusté est celui à quatre paramètres ; et l'initiation de ses paramètres se fait selon l'approche par percentiles.

3.4 Analyse de patients individuels

J'ai analysé dans un premier temps les effets des *Composé1* et *Composé2* sur l'ensemble des patients. De façon générale, les réponses d'un patient différent pour les deux composés. Cependant elles ne différent pas de la même façon d'un patient à un autre. Chez certains, la différence semble se manifester dans les réponses pour de larges concentrations, tandis que chez d'autres elle semble plutôt être au niveau des réponses pour les concentrations médianes. Pour ces premiers patients, je m'attends à avoir une différence au niveau du paramètre b lorsque les ajustements par composés seront comparés. Pour les autres patients, la différence est plus subtile, mais pourra selon moi être visible au niveau des paramètres $\log_{10} c$.

Pour vérifier ces hypothèses, j'ai aléatoirement sélectionné un patient pour chacun des profils décrits ci-haut, soit le *PatientA* pour une différence au niveau des réponses aux concentrations médianes, et le *PatientB* pour une différence au niveau des réponses pour de larges concentrations (Fig. 3.1 p.63).

Comparaison des effets de deux composés pour un patient donné. Lorsque l'on regarde les estimations des paramètres, on remarque effectivement une différence entre les valeurs de $\log_{10} c$ pour le *PatientA*, et une différence entre les valeurs de b pour le *PatientB*. Pour ce dernier patient, le paramètre s semble aussi différer d'un composé à l'autre. Avant de confirmer les hypothèses mentionnées plus haut, je suis allée examiner les intervalles de confiance des paramètres estimés ainsi que les analyses comparatives.

Nous avons démontré dans le Chapitre 2 que l'approche sélectionnée pour le calcul des

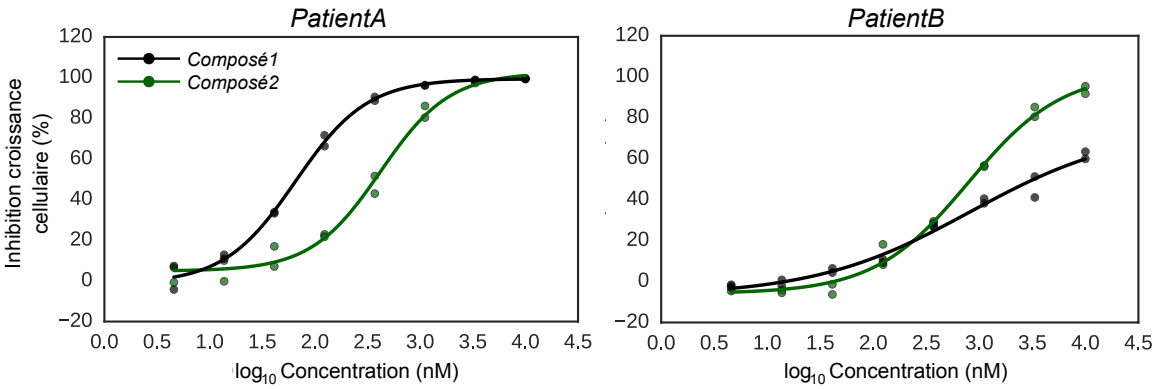


FIGURE 3.1 – AJUSTEMENTS INDIVIDUELS DES *PatientA* ET *PatientB* POUR DEUX COMPOSÉS. Les effets du *Composé1* sont représentés en noir et ceux du *Composé2* en vert. Le modèle log-logistique à quatre paramètres est ajusté individuellement selon les réponses d'un patient à un composé précis. Chaque concentration testée détient deux réplicats par patient et composé.

intervalles de la comparaison statistique dépendant du nombre de réplicats et d'observations. Les *PatientA* et *PatientB* ont chacun $R = 2$ réplicats par concentration pour un total de $N = 16$ observations par composé. N'ayant pas beaucoup de réplicats, j'ai mis de côté la SMC spécifique à chaque concentration (SMC1) et ai testé la SMC avec erreur constante (SMC2-EMQ), la SMC avec erreur constante selon les données (SMC3) ainsi que le RB sur les réplicats de chaque concentration (RBC). Les écarts-types calculés lors de la SMC3 sont très semblables aux EMQs (Tables VII & VIII p.65-67), soit 2.28 et 3.55 pour le *PatientA* (*Composé1* et *Composé2* respectivement), et de 2.17 et 2.41 pour le *PatientB*. Les intervalles générés par la SMC2-EMQ et la SMC3 sont très semblables. Pour simplifier l'analyse, je ne considère que la SMC2-EMQ et la RBC (Fig. 3.2 & 3.3 p.64-68). Les deux approches testées semblent mener vers des conclusions lègèrement différentes quant à la comparaison de l'efficacité du *Composé1* et du *Composé2* chez les deux patients.

Analyse du *PatientA*. Chez le *PatientA*, les deux approches confirment que les valeur de $\log_{10} c$ sont significativement différentes pour les deux composés (en gras dans la table VII p.65). La figure 3.2 illustre bien cette différence. Sur cette même figure, on remarque un certain motif dans les nuages de points du *Composé1* pour le RBC. Effectivement, les

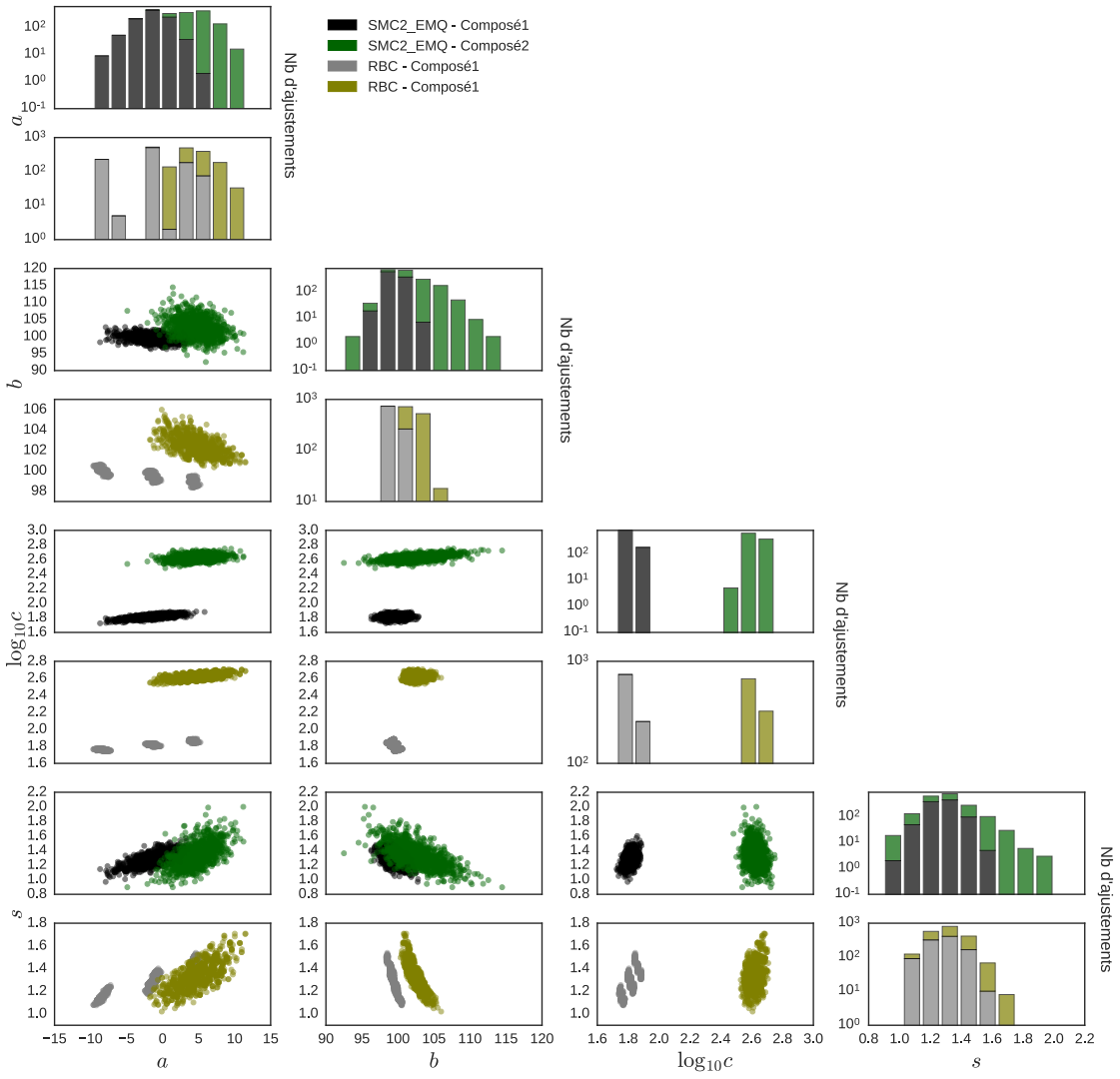


Figure 3.2 – Comparaison des paramètres estimés pour le *PatientA* par SMC2-EMQ et RBC. Les données de SMC2-EMQ sont en noir et vert, et les données de RBC sont en jaune-olive et gris. Les données du *Composé1* sont en noir et gris, et celles du *Composé2* en vert et jaune-olive. Chaque colonne est représentative d'un paramètre, tandis que chaque deux rangées sont représentatives d'un paramètre. Les histogrammes illustrent les distributions des valeurs de chaque paramètre pour 1 000 SMC2-EMQ et RBC. Les nuages de points illustrent la relation entre deux paramètres. Les p-values pour chaque paramètre et approche sont reportées dans la table VII. Le motif entrecoupé observé pour les RBC du *Composé1* est dû à la variance entre les réplicats de petite concentration (Fig. 3.1 p.63).

Table VII – Comparaison des ajustements du *PatientA* pour différents composés

	Composé1	*Composé2*	p-value SMC2-EMQ	p-value RBC
a	-1.38	4.83	**0.01**	0.1
b	99.45	102.48	0.13	<**0.00009**
$\log_{10} c$	1.82	2.62	<**0.00009**	<**0.00009**
s	1.28	1.32	0.4	0.39
EMQ	2.30	3.97		

expériences de ré-échantillonnage semblent se regrouper en trois groupes distincts pour les paramètres a, $\log_{10} c$ et s. Pour expliquer cela, retournons aux données d'inhibition de la croissance cellulaire pour ce composé (Fig. 3.1 p.63). Nous remarquons alors que les réplicats pour une même concentration sont globalement très semblables. Il semble cependant avoir une plus grande variance dans les valeurs de la plus petite concentration. Les jeux de données obtenus lors des expériences de ré-échantillonnage sont alors eux aussi très semblables et différent principalement de par leurs valeurs pour la plus petite concentration : ceux ayant deux fois le réplicat R_1, ceux ayant le réplicat R_1 et le réplicat R_2, et ceux ayant deux fois le réplicat R_2. Les valeurs pour de grandes concentrations ne différent quasiment pas d'un ré-échantillonnage à un autre. Cela résulte en un paramètre b très stable entre les ré-échantillonnages. Ce sont donc les paramètres a, $\log_{10} c$ et s qui sont principalement affectés par les valeurs de la plus petite concentration, d'où les trois regroupements de valeurs pour ces paramètres dans la figure 3.2. L'approche par RBC ne semble donc pas être appropriée à l'analyse des données du *PatientA*.

Avant d'enchaîner avec l'analyse des résultats de la SMC2-EMQ, je me permets d'utiliser les résultats de la RBC pour préciser un point important de l'analyse comparative. On remarquer que la stabilité du paramètre b pour les différents ré-échantillonnage mène à une p-value significative lorsque l'on compare les valeurs de ce paramètre pour les *Composé1* et *Composé2*. Bien que les valeurs en tant que telles se ressemblent (Table VII p.65), les *Composé1* et *Composé2* ont des réponses maximales distinctes, selon l'approche de RBC. Il est important ici de bien saisir la signification d'un tel résultat. La différence statis-

tique entre des valeurs semblables d'un paramètre est une confirmation de la fiabilité des ajustements initiaux. Dans le cas présent, conclure que le *Composé2* génère une réponse maximale plus élévée que le *Composé1* chez le *PatientA*, bien que vrai, n'est pas très informatif. Or, nous pouvons conclure avec confiance que les deux composés génèrent des réponses quasi-optimale (c'est-à-dire de 100%) chez le *PatientA*.

La SMC2-EMQ est plus *lousse* que le RBC puisque l'erreur estimée par l'EMQ est légèrement plus élevée que la variance par concentration des réplicats. Contrairement au RBC, les estimations des paramètres pour les différentes simulations ne se regroupent pas en sous-groupe, mais restent homogènes. La tendance des différents nuages de points reste cependant sensiblement la même que pour le RBC. Outre le pramètre $\log_{10} c$, la SMC2-EMQ relève une différence significative entre les paramètres a des deux ajustements (en gras dans la table VII p.65). Cela est intéressant puisque nous avons vu que les valeurs des différents paramètres étaient dépendantes : ce pourrait-il que la différence dans les valeurs du paramètre $\log_{10} c$ soit un résutlat de la différence dans les valeurs du paramètre a ? Les valeurs de a pour les ajustements initiaux ne diffèrent que d'environ 6 unités et se rapprochent toutes deux de la valeur optimale du contexte expériemental qui est de 0.00. Une aussi petite différence dans les paramètres a ne causerait pas une aussi grande différence entre les valeurs du paramètre $\log_{10} c$, surtout lorsque les valeurs des paramètres b et s sont très semblables entre les deux ajustements. la p-value obtenue pour la comparaison des paramètres $\log_{10} c$ est plus petite que celle des paramètres a. La différence entre les paramètres $\log_{10} c$ n'est donc fort probablement pas causée par la différence dans les valeurs du paramètre a.

Les résultats de l'analyse des données du *PatientA* suggèrent que bien que les deux composés aient une efficacité semblable en terme de réponse optimale ($a \approx 100.00\%$), le *Composé1* semble agir plus efficacement (c'est-à-dire à plus petite concentration) que le *Composé2* ($\log_{10} c_1 < \log_{10} c_2$). Le *Composé1* serait donc plus efficace chez le *PatientA*, selon cette expérience de criblage.

Table VIII – Comparaison des ajustements du *PatientB* pour différents composés

	Composé1	*Composé2*	p-value SMC2-EMQ	p-value RBC
a	-7.11	-5.89	0.37	0.32
b	75.41	102.63	0.12	0.08
$\log_{10} c$	2.91	2.90	0.54	0.55
s	0.59	0.98	**0.01**	$<$**0.00009**
EMQ	3.12	3.21		

Analyse du PatientB. Chez le *PatientB*, les deux approches génèrent des valeurs similaires et des conclusions semblables. Contrairement à notre hypothèse initiale, la valeur du paramètre b du *Composé2* n'est pas significativement supérieure à celle du *Composé1*. Cependant, le paramètre s du *Composé2* est significativement plus grand que celui du *Composé1*. Les deux approches, la SMC2-EMQ et le RBC, relèvent cette différence statistique (Table VIII p.67). Il semble avoir beaucoup de variation dans les valeurs estimées pour les différents paramètres lors des expériences de simulation de données et de ré-échantillonnage. Les paramètres a et b semblent avoir plusieurs valeurs extrêmes sortant du contexte biologique de l'analyse et affectant les valeurs des paramètres $\log_{10} c$ et s. Cette grande variation dans les valeurs de paramètres est principalement présente dans les données du *Composé1* (Fig. 3.3 p.68). Lorsque nous retournons aux données expérimentales initiales du *PatientB* (Fig. 3.1 p.63), nous remarquons que les données pour le *Composé1* ne semblent pas former de plateaux à de basses et hautes concentrations. Cela suggèrent que les données pourraient être insuffisantes pour estimer les paramètres avec confiance. Lors de l'ajustement, le modèle doit anticiper les réponses à des concentrations qui n'ont pas été testé expérimentalement. Cela peut alors mener à des paramètres a et b irréalistes ($a \ll 0.00$ et $b \gg 100.00$), ainsi qu'à un IC_{50} supérieur à la plus grande concentration testée expérimentalement (traits rouges dans la Fig. 3.3 p.68). Contrairement au *Composé1*, les données pour le *Composé2* semblent avoir des plateaux plus définis, ce qui explique la plus petite variation dans les estimations des paramètres lors des expériences de simulation et de ré-échantillonnage de données.

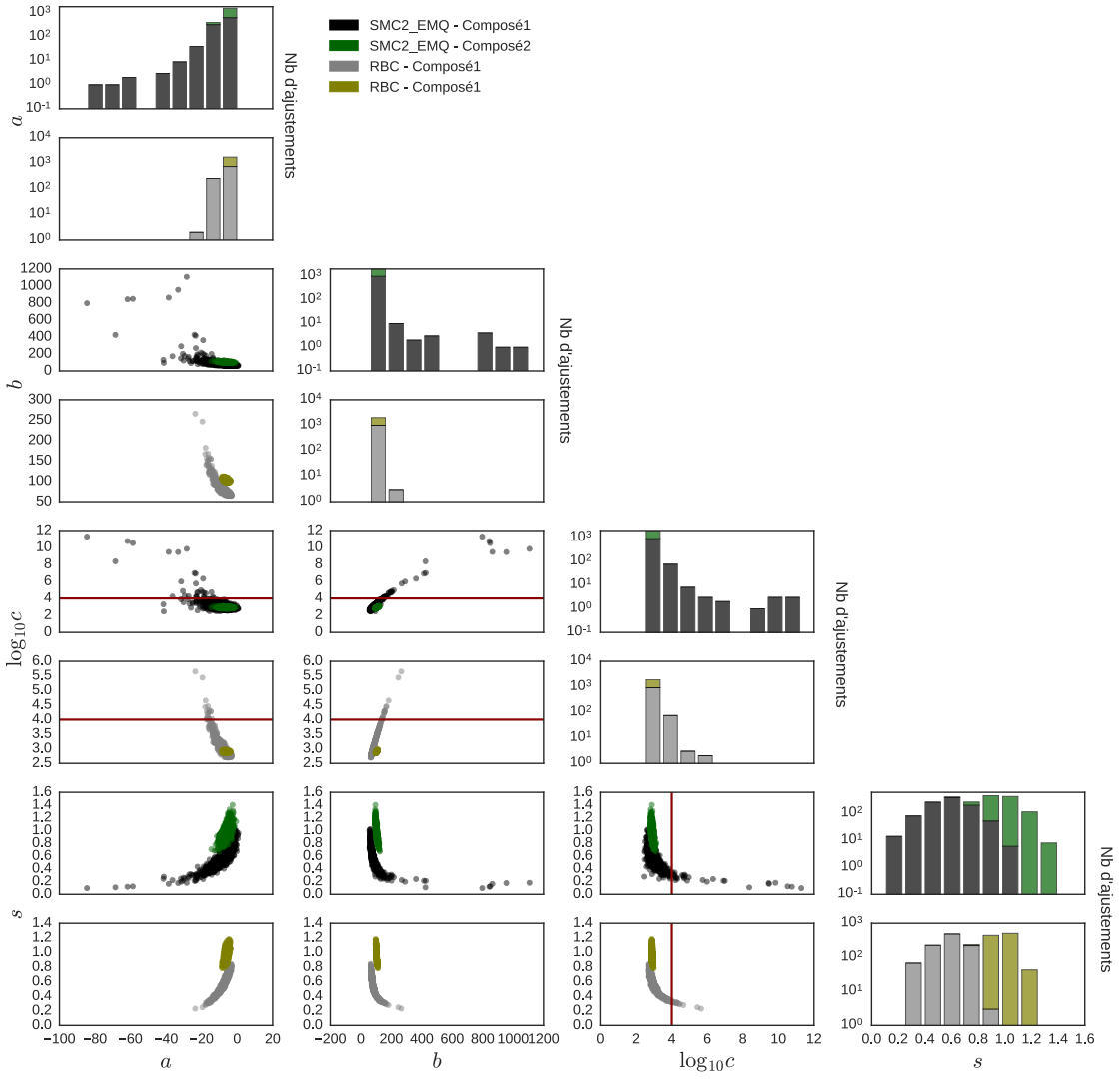


FIGURE 3.3 – COMPARAISON DES PARAMÈTRES ESTIMÉS POUR LE *PatientB* PAR SMC2-EMQ ET RBC. Les données de SMC2-EMQ sont en noir et vert, et les données de RBC sont en jaune-olive et gris. Les données du *Composé1* sont en noir et gris, et celles du *Composé2* en vert et jaune-olive. Chaque colonne est représentative d'un paramètre, tandis que chaque deux rangées sont représentatives d'un paramètre. Les histogrammes illustrent les distributions des valeurs de chaque paramètre pour 1 000 SMC2-EMQ et RBC. Les nuages de points illustrent la relation entre deux paramètres. Les segments rouges représentent la valeur de la concentration expérimentale maximale. Les p-values pour chaque paramètre et approche sont repportées dans la table VIII.

Les résultats de l'analyse des données du *PatientB* suggèrent que le *Composé2* n'est pas nécessairement plus efficace en terme de réponse maximale que le *Composé1*. Cependant, la pente de le courbe dose-réponse de ce dernier est significativement plus petite que celle du *Composé2*. Dans le contexte biologique de l'analyse, cela signifierait que le gain d'efficacité du *Composé2* est plus rapide que celui du *Composé1*. Une plus large valeur du paramètre s indique que le dosage du composé doit être précis. Dans le contexte du développement de médicament, cette information doit être prise en compte.

Pour une expérience donnée, nous remarquons que les effets des *Composé1* et *Composé2* ne sont pas les mêmes. Ils semblent aussi que ces effets soient différents d'un patient à un autre. Dans le but d'approfondir notre analyse, nous avons regardé ce qui pourrait différentier nos deux patients au niveau génétique. Plusieurs différences ont été relevées par mes collaborateurs du laboratoire Sauvageau. Ce basant sur des travaux préliminaires (non publiés), nous avons décidé d'explorer plus en profondeur une de ces différences, soit le fait que le *patientB* soit muté au niveau du gène W et que le *PatientA* ne le soit pas. Il est à noter, que l'appélation "gène W" est fictive et vise à anonymiser le gène en question. La prochaine section fait le portrait des analyses de groupes faites dans ce contexte.

3.5 Analyse de groupes de patients

Suite aux analyses individuelles des effets de deux composés sur différents patients, nous avons décidé d'appronfondir notre analyse de ces mêmes composés en regardant leurs effets sur différents groupes de patients. Le *GroupeA* est un ensemble de 66 patients portant tous une mutation au gène W. Le *GroupeB* est quant à lui l'ensemble des 115 autres patients de l'expérience. Ceux-ci ne sont pas mutés au gène W. Les composés étudiés sont les mêmes que pour les analyses par patient, soit les *Composé1* et *Composé2*.

Dans la dernière section, nous étions intéressés à savoir si pour un même patient ces composés avaient des effets différents. Pour l'analyse des groupes, nous voulons mainte-

nant savoir si un composé donné à des effets différents sur deux ensembles de patients. Les résultats obtenus à la dernière section suggèrent que pour le *Composé1* les groupes seront différents au niveau des paramètres b et $\log_{10} c$, tandis que pour *Composé2* se sera principalement au niveau du paramètre s, avec une légère différence dans les valeurs du paramètre $\log_{10} c$. Analysons les données de la figure 3.4 pour confirmer ou infirmer ces hypothèses.

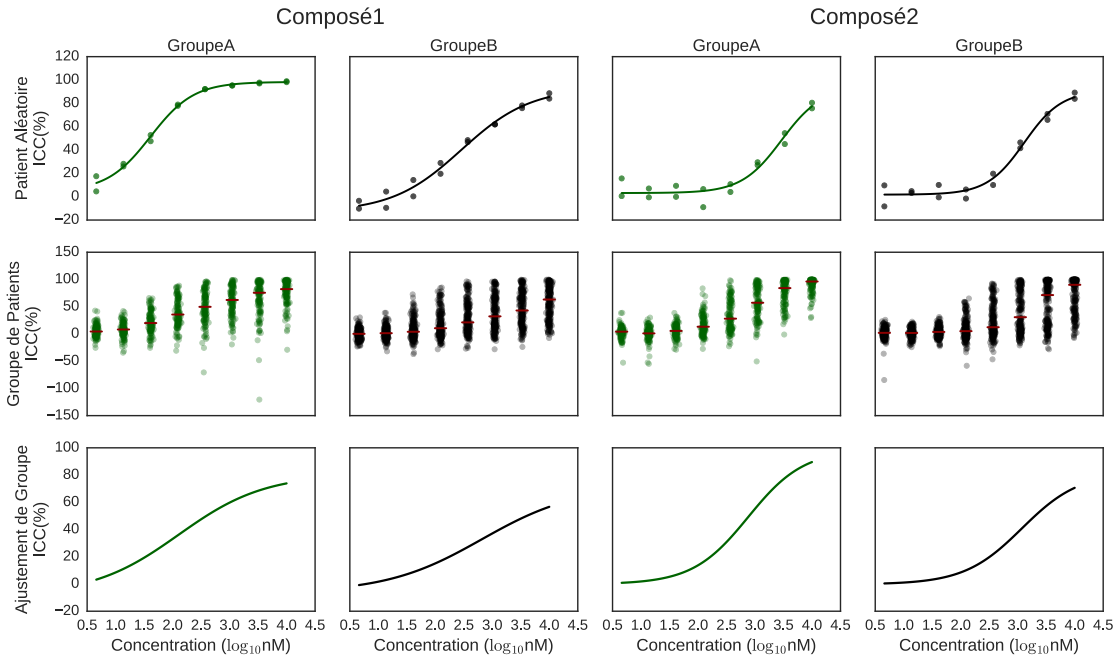


Figure 3.4 – Données et ajustements des effets des *Composé1* et *Composé2* sur deux groupes de patients. Le *GroupeA* compte 66 patients tandis que le *GroupeB* en compte 115. Les groupes sont les mêmes pour les deux composés. Première rangée : pour chaque paire groupe-composé, les réponses et l'ajustement individuel d'un patient aléatoire sont illustrés. Deuxième rangée : un bruit gaussien aléatoire ($\sigma = 0.02$) a été ajouté aux concentrations associées aux réponses de l'ensemble des patients d'un groupe données. Les tirets rouges sont représentatifs des réponses médianes par concentrations. Troisième rangée : les ajustements de groupe sont obtenus selon l'ensemble des réponses des patients pour un groupe donné. Les données du *GroupeA* sont représentées en vert et celles du *GroupeB* en noir. ICC = Inhibition de la croissance cellulaire.

Comparaison des effets d'un composé sur deux groupes de patients. Le mo-

Table IX – Estimations de paramètres pour les ajustements de groupes

	Composé1		***Composé2***	
	GroupeA	*GroupeB*	*GroupeA*	*GroupeB*
a	-8.26	-6.36	-0.16	-0.29
b	80.58	72.05	98.55	82.0
$\log_{10} c$	2.1	2.81	2.88	3.07
s	0.57	0.52	0.89	0.86
EMQ	24.02	24,24	19.94	26.37

dèle log-logistique à quatre paramètres est ajusté selon les données de chacun des groupes (Fig. 3.4 p.70). Pour un ensemble de patients nous obtenons des estimations pour les paramètres a, b,
corlog$_{10}$ c et s qui sont représentatifs de l'effet global du composé étudié pour ce groupe. La comparaison des effets d'un composé sur deux groupes de patients se fait par réechantillonnage Bootstrap sur les patients (RBP) (Fig. 3.5 p.73). J'ai choisi d'implémenter et d'utiliser cette approche pour deux raisons.

Premièrement, l'approche par simulation Monte-Carlo spécifique à chaque concentration (SMC1) ne peut être appliquée dans ce contexte. Ayant plusieurs réponses par concentration, j'ai évalué la normalité de ces données avec le test de Shapiro-Wilk [54]. Pour une valeur α donnée, nous ne pouvons rejeter l'hypothèse nulle, soit que les données proviennent d'une distribution normale, lorsque p-value $> \alpha$. Dans le cas contraire, lorsque la p-value est inférieure à α, nous pouvons conclure que les données ne provienent pas d'une distribution normale. Or, pour un $\alpha = 0.05$, les quatres ensembles de données utilisés contiennent chacun au plus deux concentrations pour lesquelles nous ne pouvons rejetter l'hypothèse nulle de normalité (en gras dans la table X p.72). L'approche de la SMC1 assume une erreur normale, ce qui dans ce contexte d'analyse ne semble pas représenter la réalité.

Deuxièmement, les approches de simulation Monte-Carlo avec erreur constante (SMC2 et SMC3) semblent elles aussi ne pas convenir à l'analyse. J'ai appliqué le test de Shapiro-

Table X – Résultats (p-value) du test Shapiro-Wilk sur les réponses par concentration

Concentration($\log_{10}$ nM)	***Composé1***		***Composé2***	
	GroupeA	*GroupeB*	*GroupeA*	*GroupeB*
0.66	0.023	**0.538**	< 0.001	< 0.001
1.14	**0.300**	**0.066**	< 0.001	**0.538**
1.61	**0.320**	< 0.001	**0.891**	**0.162**
2.09	0.044	< 0.001	0.001	< 0.001
2.57	< 0.001	< 0.001	< 0.001	< 0.001
3.04	< 0.001	< 0.001	< 0.001	< 0.001
3.52	< 0.001	< 0.001	< 0.001	< 0.001
4.00	< 0.001	< 0.001	< 0.001	< 0.001

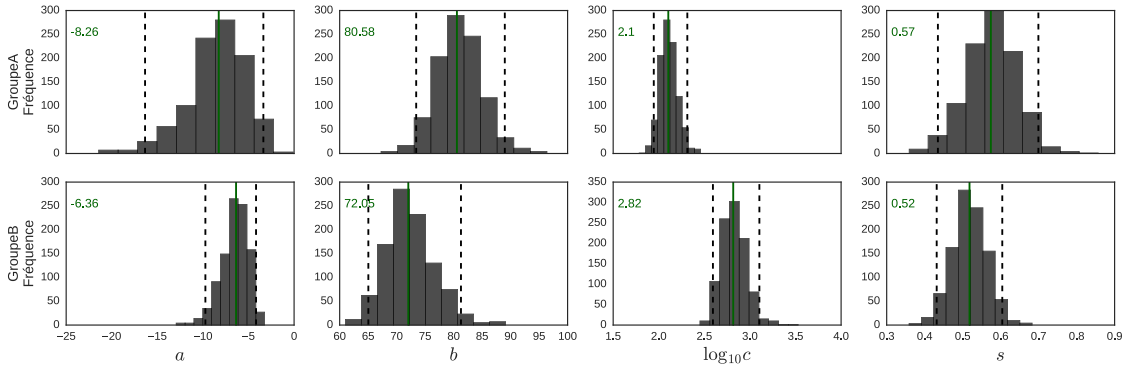
Wilk sur l'ensemble des résiduels pour chacune des paires groupe-composé. Les p-values pointaient encore vers le rejet de l'hypothèse nulle, soit 8.83×10^{-32} et 7.47×10^{-35} pour les *GroupeA* et *GroupeB* du *Composé1*, respectivement, et 1.35×10^{-29} et 9.80×10^{-33} pour le *Composé2*. De plus, je me suis demandée s'il serait valable de représenter l'erreur par une distribution normale commune à toutes les données, sachant que l'erreur par concentration ne semble pas être normale ? Selon le théroème de la limite centrale (de l'anglais *central limit theorem*), si nous avons assez de données au total, le bruit commun devrait être normal. De par nos expériences sur des données synthétiques, nous avons remarqué que l'estimation de l'erreur était grandement affectée par le nombre de réplicats présents par concentration. Dans le cas de l'analyse de groupe, les réplicats sont les patients du groupe. Il semble alors plus pertinent d'appliquer notre approche de ré-échantillonnage sur les patients mêmes plutôt que sur l'ensemble des données.

Le ré-échantillonnage Bootstrap sur les patients (RBP) nous permet d'établir la fiabilité des paramètres estimés en inférant l'erreur causée par la mise en commun des données de plusieurs patients. Cette approche nous permet aussi d'évaluer la stabilité de nos groupes, à savoir s'il y a beaucoup de variation dans les paramètres prédits pour un groupe lorsque l'on ré-échantillonne les patients qui le forme.

Composé1. Les effets du *Composé1* sur les *GroupeA* et *GroupeB* semblent différer

principalement au niveau des réponses maximales et des IC_{50}, soit au niveau des valeurs des paramètres b et $\log_{10} c$, tel qu'attendu. Les paramètres a et s ont des valeurs semblables pour les deux groupes. De façon générale, les intervalles de confiance semblent être plus larges pour les estimations du *GroupeA* (Fig. 3.5a p.73).

(a) *Composé1*



(b) *Composé2*

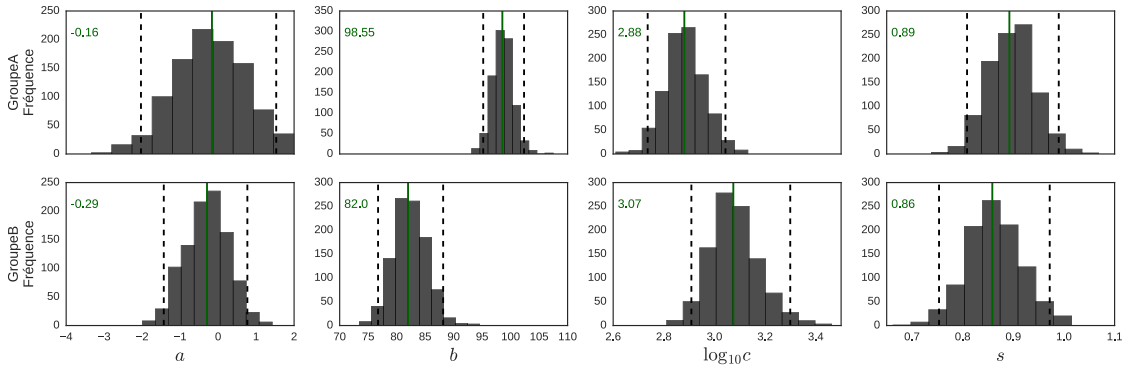


Figure 3.5 – Intervalles de confiances obtenusn par RBP pour les paramètres des *Composé1* et *Composé2*. Un histogramme est représentif de la distribution des valeurs d'un paramètre pour 1 000 RBP. Les estimations initales des paramètres sont représentées par les traits continus verts. Leurs valeurs sont aussi inscrites dans le coin supérieur gauche de chaque graphique. Les intervalles de confiance sont représentés par les traits hachés noirs.

Lorsque l'on compare statistiquement les deux ajustements avec la méthode du RBP, seule la comparaison du paramètre $\log_{10} c$ est significative : le *GroupeB* a un IC_{50} supérieur

au *GroupeA*. Il semble bel et bien avoir une différence entre les paramètres b, mais celle-ci n'est pas significative (Fig. 3.6a p.75). Il est intéressant de comparer les résultats obtenus par RBP à ceux obtenus avec le test-U Wilcoxon-Mann-Whitney (WMW) (Fig. 3.6 p.75). Les deux approches relèvent la différence dans les paramètres $\log_{10} c$. Le test-U WMW utilise les données de patients individuels pour comparer les groupes. Avec ces données, il est aussi intéressant de calculer les valeurs moyenne et médianne de chaque paramètre (Table XI p.74). Les valeurs moyennes sont grandement affectées par des patients "abérrants" ayant des paramètres extrèmes. Leurs effets se font principalement resentir sur les valeurs moyennes du paramètre b. Celles-ci sortent largement du contexte expérimental étudié. Le paramètre s est aussi affecté par de tels patients, notamment par ceux du *GroupeA*. Pour ce groupe, le paramètre s moyen est négatif ce qui inverserait le sense de notre courbe et suggèrerait donc une réponse maximale lorsqu'il y a absence de composé. Les paramètres moyens ne semblent alors ne pas être de bon indicateurs des effets d'un composé lors de l'analyse de groupe. Pour ce qui est des valeurs médianes, celles-ci se rapprochent d'avantage des valeurs estimées lors de l'ajustement de groupe sans pour autant être les mêmes (Tables X & XI p.72-74).

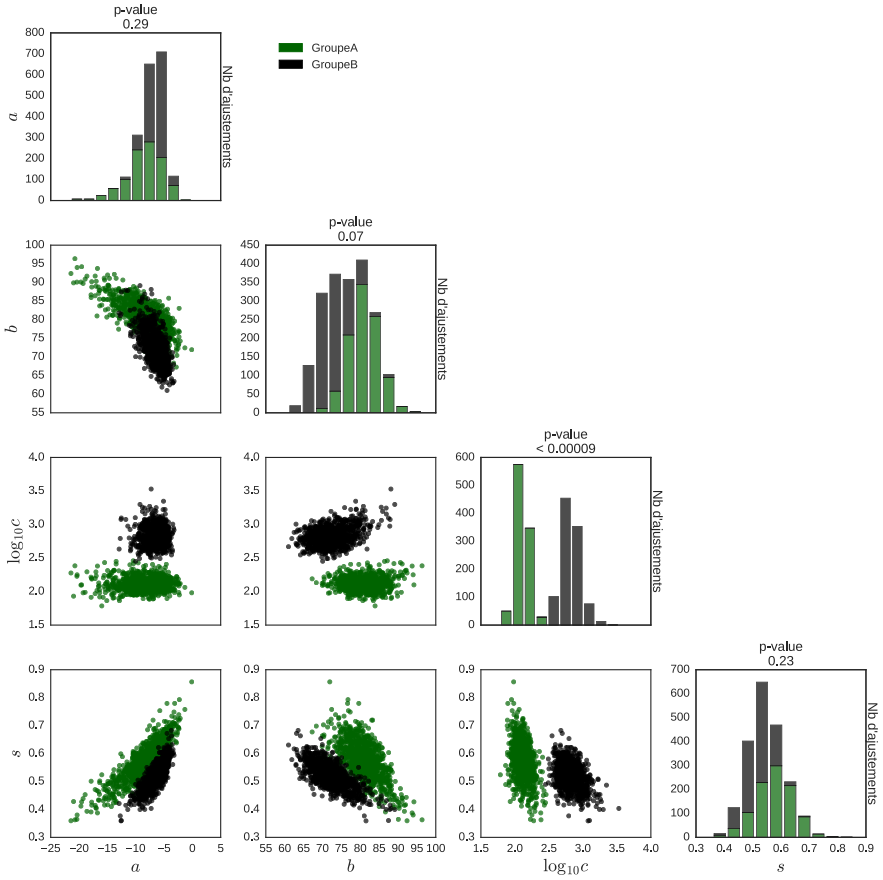
Selon les résultats présentés ci-haut, le *Composé1* serait plus efficace sur le *GroupeA* que le sur *GroupeB* : le composé semble avoir un IC_{50} statistiquement inférieur pour le premier groupe lorsque comparé à celui du deuxième groupe (Fig. 3.6 p.75).

Table XI – Valeurs moyennes et médianes des paramètres selons les ajustements par patient du *Composé1*

	GroupeA		***GroupeB***	
	Moyenne	Médiane	Moyenne	Médiane
a	-8.34	-5.16	-15.83	-3.94
b	381.85	96.14	331.95	89.16
$\log_{10} c$	2.51	2.27	3.58	3.04
s	-0.28	0.74	3.41	0.89

Composé2. Pour ce qui est des effets du *Composé2* sur les *GroupeA* et *GroupeB*, ceux-ci semblent aussi différer au niveau des réponses maximales (b) et des IC_{50} ($\log_{10} c$). Contrairement aux prédictions initiales, les valeurs du paramètre s sont très semblables

(a) Ré-échantillonnage Bootstrap sur les patients



(b) Test-U Wilcoxon-Mann-Whitney

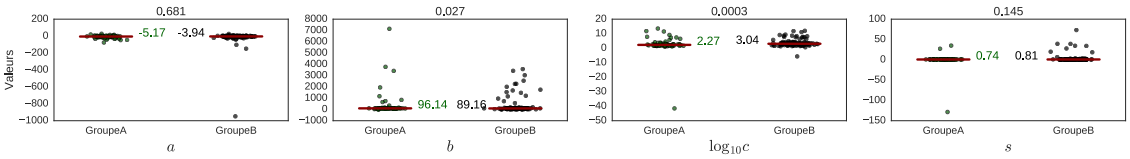


Figure 3.6 – Comparaisons des paramètres du *Composé1* pour les *GroupeA* et *GroupeB*. Les données du *GroupeA* sont représentées en vert et celles du *GroupeB* en noir. 3.6a. Comparaison des paramètres selon 1 000 RBP. Une p-value est calculée pour chaque paramètre et est inscrite au-dessus des histogrammes. 3.6b. Regroupement des paramètres de chaque patient d'un groupe. Le tiret rouge est représentatif de la valeur médiane qui est aussi inscrite à côté des nuages de points. Le test-U Wilcoxon-Mann-Whitney est fait pour comparer deux distributions ainsi que leurs médianes. La valeur du p-value est inscrite au dessus de chaque graphique.

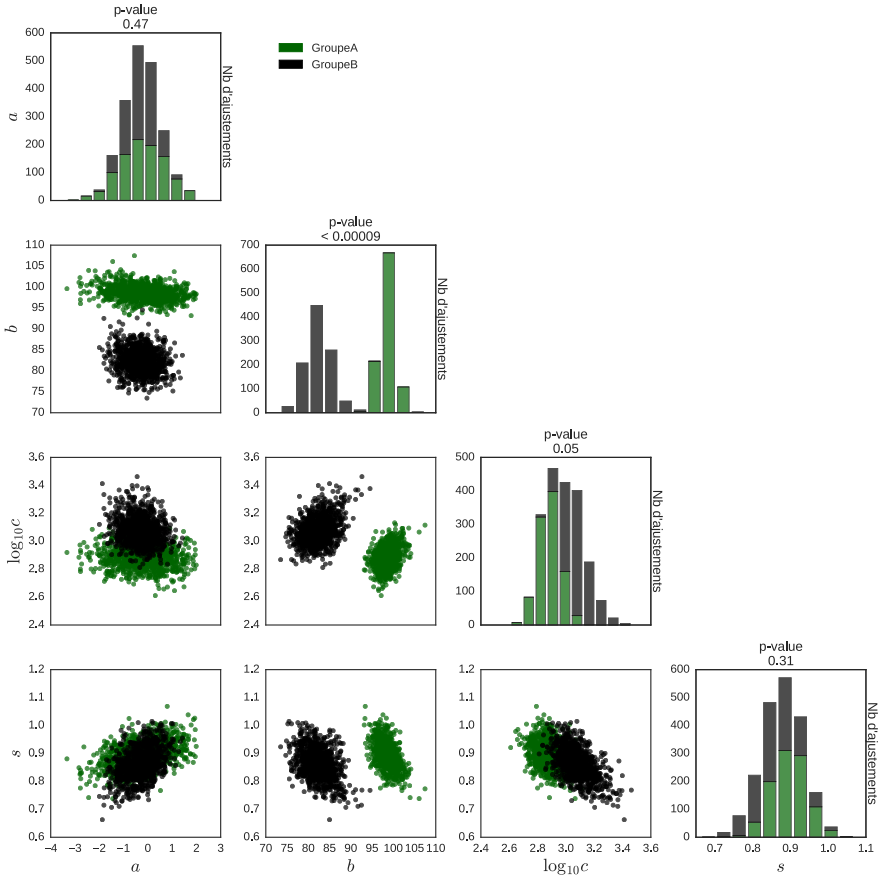
entre les ajustements. Les intervalles de confiance semblent être plus serrés pour les paramètres du *GroupeA*, à l'exception du paramètre a (Fig. 3.5b p.73).

Lorsque l'on compare statistiquement les deux ajustements avec la méthode du RBP, seule la comparaison du paramètre b est significative : le *GroupeA* semble avoir une réponse maximale supérieure à celle du *GroupeB*. Il semble aussi avoir une différence entre les paramètres $\log_{10} c$, mais celle-ci n'est pas significative (Fig. 3.7a p.77). Cela étant dit, le test-U WMW n'identifie pas le paramètre b comme étant significativement différent entre les deux groupes. De plus, ce test relève une différence entre les valeurs du paramètre $\log_{10} c$ et les valeurs du paramètre s (Fig. 3.7b p.77). Encore une fois, les valeurs moyennes des paramètres (calculées selon les valeurs obtenues lors des ajustements par patientr) ne correspondent pas aux valeurs estimées lors de l'ajustmement de groupe (Table XII p.78). Les valeurs médianes se rapprochent d'avantage des estimations faites pour les ensembles de patients, comme pour le *Composé1* (Tables X & XII p.72-78). Cependant, il y a tout de même des différences entre ces valeurs.

Le test-U WMW tente de déterminer si deux ensembles de données proviennent de la même distribution. Les résultats de ce test peuvent aussi être interprétés comme évaluant la différence entre les médianes des deux ensembles de données. Or, nos analyses indiquent une différence dans les paramètres estimés pour un groupe et les paramètres médians pour un ensemble de patients. Ce sont pas les mêmes valeurs qui sont comparées dans les deux tests. Cela pourrait expliquer les différentes conclusions obtenues. Nous discuterons d'avantage de ces deux tests dans le Chapitre 4. Nous utiliserons les résultats obtenus par RBP pour tirer nos conclusions.

Selon les résultats présentés ci-haut, le *Composé2* serait plus efficace sur le *GroupeA* que sur le *GroupeB* : le composé semble générer une réponse maximale significativement plus élevée chez le premier groupe que chez le deuxième groupe (Fig. 3.7 p.77).

(a) ré-échantillonnage Bootstrap sur les patients



(b) Wilcoxon-Mann-Whitney test U

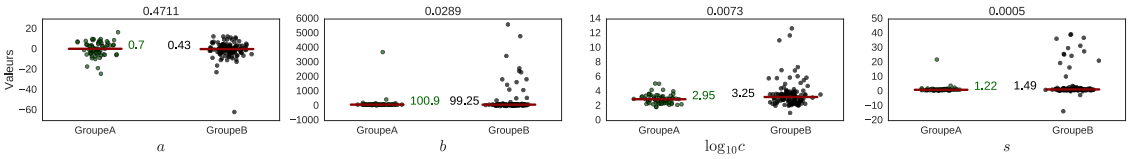


Figure 3.7 – Comparaisons des paramètres du *Composé2* pour les *GroupeA* et *GroupeB*. Les données du *GroupeA* sont représentées en vert et celles du *GroupeB* en noir. 3.7a. Comparaison des paramètres selon 1 000 RBP. Une p-value est calculée pour chaque paramètre et est inscrite au-dessus des histogrammes. 3.7b. Regroupement des paramètres de chaque patient d'un groupe. Le tiret rouge est représentatif de la valeur médiane qui est aussi inscrite à côté des nuages de points. Le test-U Wilcoxon-Mann-Whitney est fait pour comparer deux distributions ainsi que leurs médianes. La valeur du p-value est inscrite au dessus de chaque graphique.

Table XII – Valeurs moyennes et médianes des paramètres selons les ajustements par patient du *Composé2*

	GroupeA		***GroupeB***	
	Moyenne	Médiane	Moyenne	Médinne
a	0.58	0.70	-0.27	0.43
b	162.61	100.90	349.95	99.25
$\log_{10} c$	2.96	2.95	3.60	3.25
s	1.64	1.22	4.51	1.49

3.6 Conclusion

Dans ce troisième chapitre, nous avons appliqué le processus d'analyse présenté dans le Chapitre 2 à des données expérimentales. Ce travail visait à explorer les effets de deux composés chimiques sur des cellules leucémiques primaires.

Dans un premier temps, nous avons fait une analyse par patient. Pour un patient donné, nous voulions savoir si un composé semblait être plus efficace qu'un autre. L'approche statistique comparative proposée dans le Chapitre 2 fut très informative. Nous étions en mesure de déterminer pour chacun des patients étudiés quel composé semblait le plus efficace et de caractériser cette efficacité en identifiant le ou les paramètres qui différaient entre les composés.

Dans un deuxième temps, nous avons fait l'analyse de groupes de patients. Nous avons démontré que l'approche du Chapitre 2 est aussi applicable lorsque l'on compare les effets d'un composé sur différents groupes. Nous avons comparer les résultats de cette analyse à ceux obtenus lorsque l'on applique un test-U Wilcoxon-Mann-Whitney sur les valeurs des paramètres obtenues lors des ajustements individuels. Les différents résultats ne menaient pas toujours aux mêmes conclusions. Nous avons aussi relevé les différences entre les estimations des paramètres pour un ajustement de groupe et les valeurs moyennes et médianes de pour un ensemble de patients. Nous discuterons d'avantage de ces différences et de leurs implication dans le prochain chapitre.

Le Chapitre 4 abordera les différents résultats obtenus dans ce chapitre ainsi que le précédant. Une discussion sera alors faite sur l'approche proposée ainsi que les prochaines étapes à entreprendre.

Chapitre 4

Discussion

Le premier chapitre de ce mémoire présentait les multiples concepts théoriques utilisés lors de l'analyse de données de criblage à haut-débit (CHD) de type dose-réponse. Ces mêmes concepts furent mis en pratique dans le Chapitre 2 lors de la présentation du processus d'analyse proposé. Les différentes composantes de celui-ci furent présentées et plusieurs approches alternatives furent explorées pour chacune d'elles. La sélection d'une approche fut justifiée par les résultats de diverses expériences faites sur des données synthétiques. Le processus global fut par la suite utilisé lors de l'analyse de données expérimentales telles que détaillées dans le Chapitre 3.

Le présent chapitre consolidera sous forme de discussion les informations ainsi que les résultats obtenus dans les chapitres précédents. Nous discuterons entre autre de l'apport du processus proposé à l'analyse de données de CHD, du concept de *groupes*, ainsi que de futures travaux.

4.1 Le processus d'analyse proposé

L'objectif premier du présent travail était de mettre sur pied un processus automatisé, flexible et statistique pour l'analyse de données de CHD de type dose-réponse. Bien qu'il exite des outils pour de telles analyses, ceux-ci ne combinent que très rarement ces trois caractéristiques. Les outils combinant le mieux ces caractéristiques sont, selon moi, les diverses librairies Python et *packages* R. Cependant, ceux-ci demandent un important travail de mise en place ainsi qu'une certaine maîtrise des langages de programmation. Ces outils ne sont pas accessibles à tous. Les outils Prism et ActivityBase ne combinent quant à eux pas les trois caractéristiques : tous deux ne permettent pas la comparaison statistique de deux ajustements. Prims n'est pas automatisé et ActivityBase est difficilement flexible. Le processus tel que présenté dans le Chapitre 2 et tel qu'utilisé dans le Chapitre 3, est automatisé, flexible sur plusieurs aspects et statistique.

Automatisation. Le processus proposé prend en entrées des données de criblage brutes, c'est-à-dire des données de luminescence. Une normalisation automatisée des données est faite pour obtenir les taux d'inhibition de la croissance cellulaire. Les données de luminescence sont acquises automatiquement depuis une base de données. Avant de lancer le processus, des listes d'expérience, de patients ainsi que de composés doivent être soumises. Selon l'analyse choisie (par patient ou par groupe), un processus sera rouler pour chaque ensemble expérience-patient/groupe-composé. Les différents résultats seront enregistrés séparément. Il serait alors possible d'analyser tous les composés pour tous les patients d'une expérience de CHD d'un seul coup. L'automatisation du processus d'analyse est alors un grand atout lorsque l'on considère la quantité de données pouvant être générées lors d'une expérience de CHD de type dose-réponse.

Un autre important avantage de cette automatisation est l'absence de manipulations manuelles des données. Cela diminue les risques d'erreur en plus d'accélérer l'analyse. Lorsque l'on utilise Prism, les données doivent déjà être normalisées et elles doivent être entrées manuellement. Un utilisateur pourrait par erreur interchanger des valeurs

de concentrations ou de réponses ce qui affecterait les résultats et conclusions finaux.

Finalement, les données sont analysés par le même protocole. Les différentes étapes de l'analyse est alors facilement re-traçable et les résultats peuvent facilement être reproduits.

Flexible. Nous avons démontré dans le Chapitre 3 que le processus proposé peut être utilisé dans un contexte d'analyses individuelles (par patient) et d'analyses de groupe (ensemble de patients). Il n'y a aucune restriction quant à la quantité de données analysées. L'aspect d'automatisation discuté plus haut ajoute grandement à la flexibilité du processus. Il y a aussi flexibilité dans la sélection de l'approche de l'analyse statistique. Nous avons présenté dans le Chapitre 2 différentes méthodes de simulation de données ainsi que de ré-échantillonnage. Dans ce même chapitre, nous avons vu que le choix de la méthode dépend des données étudiées, principalement du nombre de réplicats par concentration. Présentement, nous devons préciser l'approche à utiliser, mais je travaille à aussi automatiser cet aspect du processus.

La flexibilité de l'approche proposée permet d'utiliser un seul processus pour différentes expériences et/ou analyses. Il y a alors uniformité entre les résultats obtenus. Cela permet aussi d'augmenter et d'assurer la traçabilité ainsi que la reproductibilité des analyses, deux concepts forts importants dans un contexte de recherche scientifique.

Statistique. Le calcul d'intervalles de confiance pour les paramètres estimés lors de l'ajustement du modèle et la comparaison statistique de deux ajustements sont des atouts importants à l'analyse et l'interprétation des données de CHD. Les intervalles permettent d'évaluer la fiabilité de nos estimations et ainsi guider notre interprétation des paramètres. Lors des analyses de groupe, les intervalles sont particulièrement utiles. En plus d'indiquer l'exactitude des estimations, ils peuvent être indicateurs de l'homogénéité de nos groupes.

Les outils couramment utilisés dont je fais mention plus haut, ne proposent pas de comparer statistiquement des ajustements. L'approche proposée dans ce travail permet cela : deux ajustements ayant un élément de contexte en commun, soit un même composé,

un même patient ou un même groupe, peuvent être comparés. La comparaison se base sur une relation de grandeur entre les valeurs estimées d'un paramètre donné. Il est alors possible de déterminer si le paramètre p d'un ajustement est statistiquement plus grand que le paramètre p d'un autre ajustement. En comparant les paramètres de façon individuelle, nous pouvons identifier précisément la différence des effets du ou des composés pour les différentes conditions étudiées. Une différence au niveau des IC_{50} ne peut être interprétée de la même façon qu'une différence au niveau des réponses maximales. Dans le contexte du développement de médicament et de thérapies personnalisées, notre approche comparative peut être très informative, tel que démontré par les résultats obtenus dans le Chapitre 3.

4.2 Le concept du groupe de patients

J'ai introduit le concept du *groupe de patients* de façon théorique dans le Chapitre 3, puis nous l'avons testé sur des données expérimentales dans le Chapitre 3. Je définis un groupe de patients comme étant la mise en commun des réponses individuelles de plusieurs patients pour former un unique jeu de données. Les groupes sont formés à la discrétion du chercheur. Par exemple, dans le Chapitre 3, nous avons diviser les patients d'une expérience de criblage en deux groupes selon s'ils étaient mutés ou non au niveau du gène W.

Je propose d'utiliser ces groupes pour modéliser les réponses globales d'un ensemble de patients à un composé donné. Bien que le concept d'analyser les effets d'un composé sur un ensemble de patients soit courant, l'usage explicite d'un groupe de patients ne semble pas l'être. Un grand nombre de publications ajustent individuellement le modèle log-logistique aux données de divers patients pour ensuite former des groupes avec les paramètres estimés [8, 38, 23, 55, 56]. Un nombre très restreint de publications utilisent plutôt un concept se rapprochant de celui du groupe de patients [11, 17, 57].

Cette approche permet de modéliser les effets généraux d'un composé sur un ensemble de patient. Nous obtenons alors une seule valeur pour chacun des paramètres a, b, c et s.

L'exactitude de ces valeurs peut être évaluée en dérivant des intervalles de confiance. Il est alors possible de savoir si la valeur d'un paramètre donné est réellement représentative de l'ensemble des patients constituant le groupe. Ces intervalles peuvent aussi être indicateurs de l'homogénéité du groupe que nous analysons. De plus, nous avons démontré dans le Chapitre 3 que les paramètres moyens et médians ne sont pas nécessairement représentifs d'un groupe. Le caclul de ces valeurs sont grandement affectés lorsqu'il y a des données aberrantes. L'ajustement des groupes ne fait pas abstraction de ces données : leurs effets sont limiter lors de l'ajustement compte tenu du grand nombre de données utilisées. Notre approche semble mieux gérer les données abbérrantes.

Un article publié par DeLean [17] propose d'analyser simultanément les sigmoïdes de composés analagues. Certains paramètres seraient communs à tous les composés et leur ajustement se ferait en considérant toutes les réponses. Les autres paramètres seraient quant à eux ajustés individuellement aux données de chaque composé. L'approche que nous proposons ressemble à celle de DeLean : plutôt que d'analyser simultanément des composés, nous analysons simultanément les réponses de plusieurs patients. J'ai exploré lors de travaux préliminaires (non-présentés dans ce mémoire) le partage de quelques paramètres. Pour certains paramètres nous obtenions une seule valeur, tandis que pour d'autres nous avions autant de valeurs que de patients. L'analyse de ces résultats était complexe et revenait en grande partie à combiner les paramètres d'ajustement individuel. J'ai alors décidé que tous les paramètres seraient communs pour tous les patients d'un groupe, et qu'il y aurait ainsi qu'une seule sigmoïde (ajustement).

L'analyse des données par groupe de patient telle que présentée dans ce travail serait bénéfique aux recherches entreprises dans le contexte du développement de thérapie personnalisée. Il semble plus efficace d'interpréter et d'analyser les métriques pour un groupe donnée, plutôt que les métriques moyennes ou médiannes. Bien que certaines études ont déjà utilisées une telle approche semblable à la nôtre [11, 17, 57], il n'est pas facile de faire de telles analyses avec les outils mentionnés plus haut. Cela étant dit, l'analyse de groupe peut se faire très facilement avec le processus automatisé présenté dans le Chapitre 2.

4.3 L'interprétation des ajustements

Plus souvent qu'autrement, l'efficacité d'un composé chimique est caractérisé par son IC_{50} uniquement [58, 59, 60, 61, 62, 56, 8]. Bien que l'IC_{50} soit une composante importante de l'efficacité d'un composé, des études démontrent l'importance de considérer l'effet maximale ainsi que la pente de la courbe dose-réponse [63, 23]. Ces deux dernières composantes de l'efficacité sont équivalentes aux paramètres b et s, respectivement, du modèle log-logistique.

Prenons le cas hypothétique de deux composés ayant des IC_{50} égaux. Nous pourrions conclure qu'aucun des composés n'est plus efficace que l'autre. Cependant, lorsque l'on analyse l'effet maximal de ces composés, on remarque une valeur de 100% pour l'un et de 50% pour l'autre. Notre conclusion initiale ne tient plus puisqu'un taux d'inhibition de la croissance cellulaire égale à 100% est nettement supérieur à un taux de 50%. Un des composé est plus efficace que l'autre. s

Dans cet exemple, l'efficacité des composés n'est pas déterminеée par une variation dans le paramètre c, mais plutôt par une variation du paramètre b. Bien que les trois métriques mentionnées plus haut soient toutes représentatives de l'efficacité d'un composé, elles détiennent des interprétations bien distinctes, nous informant sur différents aspects du composé. L'IC_{50} est représentatif d'une concentration. Il peut nous informer quant à la puissance d'un composé. La réponse maximale est informative de l'efficience du composé : elle est représentative de ce que le composé à de mieux à nous offrir. Finalement, la pente nous informe sur la rapidité des effets du composé. Cette dernière est très importante dans le contexte du développement de médicament, car elle indique la vitesse de transition entre les réponses plus faibles et les réponses plus importantes. Dans le cas où la réponse maximale n'est pas souhaitable, le dosage du composé doit être très précis lorsque la pente est abrupte.

Quelques récentes études suggèrent une relation entre les paramètres du modèle log-

logistique, le type cellulaire ainsi que les classes de composés [63, 64, 65]. Certains résultats démontrent que la variation dans les valeurs des réponses maximales et des pentes serait assossiée aux classes de composés [63]. Analyser et interpréter ces métriques pourraient être informateur quant au mode d'action d'un composé.

Nous avons mentionner dans le Chapitre 2 que l'aire sous la courbe (AUC) dose-réponse était parfois utiliser lors de l'analyse [56, 42, 43, 41, 58]. Cependant, nous n'avons pas utilisé cette métrique lors de l'analyse des données expérimentales du Chapitre 3. Le but de ces analyses étaient de relever des différences dans les effets de deux composés pour divers contextes. Selon moi, l'AUC n'est pas une métrique idéale pour ce genre d'analyse. Premièrement, tel que mentionné dans le Chapitre 2, la valeur ainsi que les unités de l'AUC ne sont pas triviales à interpréter. Bien que l'on puisse comparer les valeurs d'AUC pour différents ajustements, les conclusions ne sont pas aussi précises que pour les autres paramètres. Deuxièmement, pour bien analyser les réponses et bien comprendre les effets du composé, il faudrait comparer l'AUC aux valeurs des paramètres b et c. L'AUC est représentative de la combinaison des effets de ces deux paramètres sur les réponses. Or, si l'on compare deux ajustements, comment savoir lequel des paramètres diffèrent (ou même savoir si les deux paramètres diffèrent) en ne comparant que l'AUC ? Il faudrait donc en plus de comparer l'AUC comparer les valeurs pour ces deux paramètres. Il semble plus simple d'analyser ces paramètres tel quel plutôt que de décortiquer leur combinaison. Finalement, l'AUC est très dépendante du protocole expérimental. Puisqu'elle est normalement calculer pour une courbe allant de la concentration minimale à la concentration maximale, différentes ensembles de concentration génèreront différentes AUCs. Il est peu souhaitable de basée une analyse sur une métrique qui dépend autant du protocole expérimental.

4.4 Les intervalles de confiance

Un intervalle de confiance permet d'éavaluer la fiabilité d'une valeur. Dans le contexte expérimental de ce travail, l'intervalle de confiance sur les estimations des paramètres

nous indique à quel point ces valeurs sont représentatives du jeu de données analysé. Nous utilisons deux approches pour calculer ces intervalles de confiance, soit la simulation de données Monte-Carlo [27] et le ré-échantillonnage Bootstrap [66]. Ces deux approches consistes à générer plusieurs nouveaux jeux de données et ensuite obtenir plusieurs nouvelles estimations de paramètres du modèle log-logistique. Les nouveaux jeux de données sont construits selon l'information contenue dans le jeu initial. De la sorte, la variation dans les valeurs des différents paramètres sont représentatifs de l'exactitude des paramètres initiaux. Il n'est cependant pas simple de déterminer la *meilleure* façon de générer ces nouveaux jeux de données, c'est-à-dire de savoir laquelle des deux approches est la plus adéquate.

L'approche par simulation de données Monte-Carlo nécessite de connaître ou d'assumer la distribution de l'erreur. Puisque la fonction objective utilisée, celle des moindres carrés, assume que l'erreur expérimentale est normalement distribuée et constante pour toutes les observations (voir Chapitre 1), nous utilisons une distribution normale pour simuler les noubelles données. Cependant, nous avons remarqué dans le Chapitre 3s que cela ne semble pas toujours être un le cas. De plus, pour de petits jeux de données, il est difficile d'estimer adéquatement l'écart-type de la distribution normale. L'erreur moyenne quadratique (EMQ) semble être une bonne estimation de l'écart-type. L'outil d'ajustement de courbe de Prism GraphPad utilise la simulation de Monte-Carlo pour générer ces intervalles de confiance. La distribution choisie ainsi que le pourquoi de cette décision ne sont cependant pas clairement expliqués.

Le ré-échantillonnage Bootstrap est alors une bonne alternative. Cette approche ne nécessite aucune précision quant à la distribution de l'erreur. Le Bootstrap détermine de façon indirecte la distribution du bruit et est très utile lorsque nous avons peu de données. De plus, cette approche permet d'évaluer la stabilité des données analysées, un aspect que la simulation de Monte-Carlo ne fait pas [67]. Cependant, le Bootstrap à tendance à être plus optimiste que la simulation de Monte-Carlo lorsque les données ne sont que très peu bruitées.

Les résultats du Chapitre 3 semblent indiquer que la simulation Monte-Carlo est plus appropriée à l'analyse par patient et que le ré-échantillonnage Bootstrap est plus approprié aux analyses de groupe. Pour faciliter l'utilisation du processus d'analyse proposé, nous utiliserons ces approches par défaut, en fonction de l'analyse souhaitée.

Un parallèle intéressant peut être fait avec le ré-échantillonnage Bootstrap et la validation croisée (de l'anglais *coss-validation*). Les deux approches consistent à générer de nouveaux jeux de données en échantillonant ou ré-échantillant depuis le jeu initial. La validation croisée est utilisée en apprentissage automatisé (de l'anglais *machine learning*) pour évaluer la fiabilité d'un modèle [68]. Elle consiste à entraîner ledit modèle sur un échantillon de données et de par la suite valider le modèle entraîné sur le reste des données. Cela est répété plusieurs fois pour différents échantillonnages. La fiabilité du modèle est alors représentée par la moyenne des scores obtenus lors de chaque validation. De plus, l'entraînement consiste à ajuster, par régression, les paramètres du modèle de tel sorte à être représentatif des données [69]. Ce travail se rapproche énormément de la régression non-linéraire faite dans le contexte de l'analyse de données de CHD. Serait-il alors possible d'appliquer la méthode de la validation croisée à notre processus d'analyse ? Notre contexte d'analyse comporte deux principales contraintes à cela. Premièrement, la quantité de données. Il serait difficile et peu optimal d'appliquer la validation croisée lorsque nous analysons des données par patient. Tout comme pour l'approche par ré-échantillonnage Bootstrap, il devrait avoir au minimum une réponse par concentration testée pour les échantillons d'entraînement, sans quoi le modèle ne pourra pas être ajusté adéquatement. En l'absence de réplicat par concentration, l'approche par validation croisée est alors impossible. Pour les analyses de groupe, l'approche pourrait être appliquée sur les patients : le modèle est ajusté selon les données de certains patients, puis ensuite testé sur un autre ensemble de patients. Survient alors la deuxième constrainte, soit celle du score. Sur quelle valeur se fonde la fiabilité du modèle ? Quelle métrique évaluons-nous lorsque le modèle ajusté est testé ? Bien que l'EMQ soit généralement utilisée pour évaluer un ajustement, il serait difficile d'interpréter la moyenne des EMQs lorsqu'un modèle ajusté est testé sur un sous-ensemble de patients. Cela étant dit, il serait intéressant dans de futurs travaux

d'expérimenter et d'ajuster la validation croisée au contexte d'analyse de données de CHD. Cette approche pourrait être utile lors de l'analyse de groupes de patients.

4.5 La comparaison de deux ajustements

L'interprétation d'un ajustement et l'évaluation des intervalles de confiance sont deux importantes composantes de l'analyse de données de criblage. Une autre composante importante est la comparaison. À notre connaissance, aucun outil d'ajustement de courbe ne permet de comparer deux ajustements selon leurs paramètres. Dans le Chapitre 2, nous proposons une approche pour faire de telles comparaisons.

La comparaison proposée établit si la valeur d'un paramètre donné est significativement supérieure à celle d'un autre ajustement. Pour $p_1 > p_2$, une p-value est calculée en évaluant la fréquence de $p_1 < p_2$ pour l'ensemble des simulations Monte-Carlo ou des ré-échantillonnage Bootstrap. Une p-value de 0.1 signifiie que $p_1 < p_2$ est observé dans 10% des simulations/ré-échantillonnage. Lorsque tous les paramètres sont comparés pour deux ajustements, certaines conclusions quant aux effets d'un ou des composés peuvent être proposées. Avec cette approche, il est possible de comparer les effets de deux composés sur un patient/groupe ou d'un composé sur différents patients/groupes. De plus, il nous ait possible de comparer des ajustements par patient et par groupe de patients.

Généralement, la comparaison de deux groupes de patients se fait en appliquant le test-U Wilcoxon-Mann-Whitney (WMW) [70] sur deux ensembles de valeurs pour un paramètre donnée [63]. Les valeurs proviennent des ajustements individuels par patient. Le résulat du test détermine si la distributions des valeurs d'un ensemble est stochastiquement plus grande que celle de l'autre ensemble. Une p-value significative indiquerait que les valeurs médianes des deux ensembles sont différentes. Ce test ne peut être appliquer pour la comparaison de patient individuel : les groupes doivent être formés d'au moins trois patients [71].

L'approche proposée et le test-U WMW ne sont pas équivalents : leurs résultats ne peuvent pas être interprétés de la même façon. Bien que les deux approches peuvent mener vers des conclusions semblables (Fig. 3.6 p.75), il se peut aussi que leur résultat diffèrent (Fig. 3.7 p.77). Cette différence pourrait être expliqué par la façon dont les groupes de patients sont formés. Notre approche assemble les groupes avant la paramétrisation. Les effets des données aberrantes sont alors moindre, compte tenu du nombre total de données utilisées pour l'ajustement. Les paramètres estimés sont alors plus stables. L'approche du test-U WMW assemble les groupes après la paramètrisation individuelle. Cependant, les données sont parfois insuffisantes pour bien estimer les paramètres et ce qui peut générer des valeurs aberrantes. Le test-U semble être affecté par la présence de ces paramètres aberrants et il conclut que les valeurs des deux groupes proviennent de différentes distributions (Fig. 3.7b p.77). L'approche proposée semble mieux gérer les données aberrantes que le test-U WMW. Je crois cependant que pour obtenir une analyse complète, il est mieux de considérer les résultats des deux tests.

4.6 La normalisation des données de luminescence

Dans la section 3.2 nous abordons très brièvement la normalisation des données de luminescence. La normalisation consiste à ré-échelonner la luminescence entre 0 et 100. Les valeurs normalisées sont alors des taux d'inhibition de la croissance cellulaire. Ces taux sont plus simple à interpréter qu'une valeur de luminescence. Cette approche de normalisation est largement répandue [8].

Cela étant dis, les protocoles expérimentaux utilisés pour générer les données du Chapitre 3 ne comportaient que des contrôles négatifs (cellules en l'absence de composé). Nous ne détenions donc aucun contrôle positive pour estimé les effets maximaux. Cela à deux principaux effets quant à l'analyse des données et leur normalisation.

Premièrement, nous limitons indirectement l'inhibition à 100%. De façon générale, le

ré-échelonnement des données de luminescence (l_i) en données d'inhibition (y_i) se fait selon les valeurs moyennes des contrôles positifs ($\overline{\text{ctrl}_+}$) et des contrôles négatifs ($\overline{\text{ctrl}_-}$) (Éq. 4.1).

$$y_i = 100 \times \frac{\overline{\text{ctrl}_+} - l_i}{\overline{\text{ctrl}_+} - \overline{\text{ctrl}_-}} \tag{4.1}$$

En l'absence de contrôles positifs, seuls les contrôles négatifs sont utilisés et l'inhibition maximale est estimée à 100%, soit la valeur théorique (Éq. 3.1). Dès lors, un imposons une limite supérieur, sans pour autant imposer une limite inférieure. Les taux d'inhibition de la croissance seront alors toujours inférieur à 100%. Ils peuvent cependant être sous la barre du 0%, bien que cette valeur soit la limite inférieure théorique. Ces effets sont observables sur les données de la section 3.5.

Deuxièmement, cela à pour effet de changer la distribution de l'erreur sur les données, puisque le calcul de la normalisation est un ratio. Les données de luminescence du Chapitre 3 sont divisées par la moyenne des valeurs de 28 contrôles négatifs. La variance de cette valeur moyenne est alors beaucoup plus petite que celle de chaque contrôle négatif individuel, et son effet sur la valeur du ratio est donc minime. De plus, dans le cas des données du Chapitre 3, la moyenne des contrôles négatifs est spécifique à chaque plaque d'échantillons et par le fait même, à chaque patient de l'étude. Les réponses d'un patient à un composé sont donc toutes normalisées par la même valeur contrôle qui peut être alors vue comme une constante. La normalisation a pour effet de ré-échelonner et d'inverser la distribution des réponses par patient pour un composé données.

Dans le cas des analyses de groupe, l'effet de la normalisation sur l'erreur est plus important car la moyenne des contrôles négatifs diffèrent d'un patient à l'autre. Les distributions d'erreur pré-normalisation et post-normalisation peuvent alors être grandement différentes. Cela étant dit, il est intéressant de remarqué que ces effets de la normalisation

sont généralement et couramment ignorés dans le domaine [72].

Pour contrer ces deux principaux effets de la normalisation, il serait possible d'analyser directement les données de luminescence et ainsi ignorer l'étape de normalisation. Bien que l'interprétation des paramètres résultants puisse être moins intuitive, nous n'imposons pas de limites aux données et l'erreur sur les données n'est pas altérée. Une autre alternative serait l'analyse des données par inférence bayésienne. Cette approche faciliterait entre autre la modélisation de l'erreur. J'aborde d'avantage cette seconde alternative dans la prochaine section.

4.7 Conclusion

Le présent mémoire décrit une méthodologie pour l'analyse de données de criblage. Un processus automatisé, flexible et statistique est mis sur pieds. L'automatisation et la flexibilité permettent de faire des analyses dites de groupe. Plutôt que d'analyser individuellement des patients et d'ensuite évaluer les valeurs médianes des paramètres, nous proposons de regrouper les patients avant la paramétrisation. Cette approche semble suggérer une meilleure gestion des données aberrantes. L'aspect statistique du processus comprend le calcul d'intervalles de confiance et la comparaison de deux ajustements. Les intervalles indiquent la faibilité des paramètres estimés et sont essentielle à une bonne interprétation. La comparaison permet de déterminer si la valeur d'un paramètre d'un ajustement est significativement plus élevée que celle d'un autre ajustement. Ce test est très informatif lors des analyses de groupe : nous pouvons déterminer si un composé semble plus efficace sur un certain groupe de patients, en plus d'identifier l'aspect qui différencie les deux groupes. Ce dernier point est fort intéressant et peut nous informer quant au mode d'action du composé.

L'approche statistique du processus est implémentée en utilisant les données de simulations Monte-Carlo et ré-échantillonnage Bootstrap. Trois approches ont été cernées comme

étant les plus représentatives du contexte de CHD, soit la SMC2-EMQ, le RBC et le RBP. Bien que la SMC2-EMQ soit plus précise que les approches Bootstrap, elle nécessite la présomption que l'erreur expérimentale est normale. Or, nous avons démontré que cela ne semble pour toujours être le cas. Le Bootstrap est alors une alternative fiable. Je compte automatiser le choix de l'approche pour faciliter d'avantage l'analyse.

Il serait aussi intéressant d'implémenter une méthode de normalisation permettant d'agréger les données de plusieurs différentes expériences de criblage. Avec le processus actuel, cela n'est pas recommander puisque la variation dans les données normalisées ne mèneraient à aucune conclusions concrètes. Selon moi, nous gagnerions à pouvoir combiner et comparer les données de différentes expériences. Il serait possible de créer et d'analyser plusieurs différents groupes de patients.

Je souhaite aussi explorer l'inférence bayésienne pour la prédiction de paramètres. Plutôt que de retourner une valeur et son intervalle de confiance, pourquoi ne pas retourner une distribution par paramètre ? Cela serait particulièrement utile lorsque les données ne semblent pas être suffisantes pour estimer des réponses minimale et/ou maximale. Dans de tels cas nous obtenons présentement des valeurs aberrantes qui sortent du contexte expérimental (eg. $b = 350$). Les ajustements sont alors généralement ignorés. La combinaison d'une approche par inférence bayésienne et l'utilisation de distributions *a priori* (de l'anglais *prior distributions*) pourrait nous permettent d'analyser ces jeux de données. De plus, nous avons vu plus haut que le choix entre la simulation Monte-Carlo et le ré-échantillonnage bootstrap n'est pas toujours trivial lors d'analyses comparatives. L'inférence bayésienne ne nécessite aucune de ces approches. Les paramètres de différents ajustements pourront être, en autre, comparés par le test du rapport de vraisemblance (de l'anglais *likelihood-ratio test*). Dans un autre ordre d'idées, les résultats obtenus dans le présent travail suggèrent que les réponses analysées ne sont pas distribuées normalement. Bien que l'utilisation de l'approche des moindres carrés soit pratique courante lors de l'analyse de données de CHD, cette approche assume que les données soient distribuées normalement. Les estimations de paramètres pourraient dès lors être biaisées. L'inférence

bayésienne serait aussi une façon de contourner ce problème, et faciliterait la modélisation de l'erreur sur les réponses analysées. Il serait aussi intéressant et souhaitable d'élargir notre processus à l'analyse de criblage combinatoire. L'inférence bayésienne pourrait aussi nous aidé à modéliser les effets de la combinaison de deux composés chimiques.

Finalement, il est important d'utiliser une bonne méthodologie pour l'analyse de données de criblage à haut débit. Celle-ci doit être capable d'analyser rapidement et précisément des composés pour plusieurs contextes et protocoles expérimentaux. La comparaison des effets de différents composés sur différents groupes de patients est hautement bénéfiques à la recherche portant sur le développement de nouveaux médicaments. De plus en plus d'efforts sont mis dans le développement de thérapies personnalisée et je crois que notre approche automatisée, flexible et stastique aidera à l'avancement de tels efforts.

Bibliographie

[1] K. Rudin, *Imaging in Drug Discovery and early Clinical Trials*, vol. 62. Springer Science & Business Media, 2006.

[2] P. Szymański, M. Markowicz, and E. Mikiciuk-Olasik, "Adaptation of high-throughput screening in drug discovery : toxicological screening tests," *International journal of molecular sciences*, vol. 13, no. 1, pp. 427–452, 2011.

[3] Editorial, "The academic pursuit of screening," *Nat Chem Biol*, vol. 3, no. 8, p. 433, 2007.

[4] K. H. Bleicher, H. Böhm, K. Müller, and A. I. Alanine, "A guide to drug discovery : Hit and lead generation : beyond high-throughput screening," *Nat Rev Drug Discov*, vol. 2, no. 5, p. 1086, 2003.

[5] E. Martis, R. Radhakrishnan, and R. Badve, "High-throughput screening : the hits and leads of drug discovery-an overview.," *Journal of Applied Pharmaceutical Science*, vol. 1, no. 1, pp. 2–10, 2011.

[6] A. Carnero, "High throughput screening in drug discovery," *Clin Transl Oncol*, vol. 8, no. 7, pp. 482–490, 2006.

[7] B. Chen, O. Litvin, L. Ungar, and D. Pe'er, "Context sensitive modeling of cancer drug sensitivity," *Plos One*, vol. 10, no. 8, p. e0133850, 2015.

[8] C. Pabst, J. Krosl, I. Fares, G. Boucher, *et al.*, "Identification of small molecules that support human leukemia stem cell activity ex vivo.," *Nat Methods*, vol. 11, no. 4, pp. 436–42, 2014.

[9] R. R. Neubig, M. Spedding, T. Kenakin, and A. Christopoulos, "International union of pharmacology committee on receptor nomenclature and drug classification. xxxviii. update on terms and symbols in quantitative pharmacology," *Pharmacological Reviews*, vol. 55, no. 4, pp. 597–606, 2003.

[10] D. Rodbard, "Statistical quality control and routine data processing for radioimmunoassays and immunoradiometric assays," *Clinical chemistry*, vol. 20, no. 10, pp. 1255–1270, 1974.

[11] S. S. Seefeldt, J. E. Jensen, and E. P. Fuerst, "Log-logistic analysis of herbicide dose-response relationships," *Weed technology*, vol. 9, no. 2, pp. 218–227, 1995.

[12] S. Z. Knezevic, J. C. Streibig, and C. Ritz, "Utilizing r software package for dose-response studies : the concept and data analysis," *Weed Technology*, vol. 21, no. 3, pp. 840–848, 2007.

[13] A. Marshall, *Principles of Economics : An introductory volume*. Macmillan London, 9 ed., 1961.

[14] C. Ritz, "Toward a unified approach to dose–response modeling in ecotoxicology," *Environ Toxicol Chem*, vol. 29, no. 1, pp. 220–229, 2010.

[15] D. Finney, "Radioligand assay," *Biometrics*, vol. 32, no. 4, pp. 721 – 740, 1976.

[16] J. Berkson, "Why i prefer logits to probits," *Biometrics*, vol. 7, no. 4, pp. 327–229, 1951.

[17] A. Delean, P. Munson, and D. Rodbard, "Simultaneous analysis of families of sigmoidal curves : application to bioassay, radioligand assay, and physiological dose-response curves.," *Am. J. Physiol.*, vol. 235, no. 2, pp. E97–102, 1978.

[18] J. Rawlings and W. Cure, "The weibull function as a dose-response model to describe ozone effects on crop yields 1," *Crop Science*, vol. 25, no. 5, pp. 807–814, 1985.

[19] Z. Zhang, "Parametric regression model for survival data : Weibull regression model as an example.," *Ann Transl Med*, vol. 4, no. 24, p. 484, 2016.

[20] P. Brain and R. Cousens, "An equation to describe dose responses where there is stimulation of growth at low doses," *Weed Research*, vol. 29, no. 2, pp. 93 – 96, 1989.

[21] E. J. Calabrese, "Hormesis : changing view of the dose-response, a personal account of the history and current status," *Mutation Research/Reviews in Mutation Research*, vol. 511, no. 3, pp. 181 – 189, 2002.

[22] W. N. Beckon, C. Parkins, A. Maximovich, and A. V. Beckon, "A general approach to modeling biphasic relationships," *Environ. Sci. Technol.*, vol. 42, no. 4, pp. 1308 – 1314, 2008.

[23] M. Hafner, M. Niepel, M. Chung, and P. K. Sorger, "Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs," *Nature methods*, vol. 13, no. 6, pp. 521–527, 2016.

[24] J. S. Witte and G. Sander, "A a nested approach to evaluating dose-response and trend," *Ann Epidemiol*, vol. 7, no. 3, pp. 188–193, 1997.

[25] T. P. J. Knowles, C. A. Waudby, G. L. Devlin, S. I. A. Cohen, *et al.*, "An analytical solution to the kinetics of breakable filament assembly," *Science*, vol. 326, no. 5959, pp. 1533–1537, 2009.

[26] W. W. Focke, I. van der Westhuizen, N. Musee, and M. T. Loots, "Kinetic interpretation of log-logistic dose-time response curves," *Scientific Reports*, vol. 7, no. 1, pp. 2234–2245, 2017.

[27] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in Fortran 77 : the art of scientific computing*, vol. 2. Cambridge university press Cambridge, 1992.

[28] J. O. Rawlings, S. G. Pantula, and D. A. Dickey, *Regression Diagnostic*, p. 660. Springer, 2 ed., 2001.

[29] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming : Theory and Algorithms*. John Wiley & Sons, 3 ed., 2014.

[30] J.-y. Fan, "A modified levenberg-marquardt algorithm for singular system of nonlinear equations," *Journal of Computational Mathematics*, vol. 21, no. 5, pp. 625–636, 2003.

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–538, 1986.

[32] A. Bjorck, *Numerical methods for least squares problems*. Siam, 1996.

[33] H. Yu and B. M. Wilamowski, *Levenberg-Marquardt Training*, vol. 3, p. 4052. CRC Press, 2 ed., 2011.

[34] J. Nocedal and S. J. Wright, *Sequential quadratic programming*, p. 664. Springer New York, 2 ed., 2006.

[35] M. Hagan and M. Menhaj, "Training feedforward networks with the marquardt algorithm," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989 – 993, 1994.

[36] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quarterly of applied mathematics*, vol. 2, no. 2, pp. 164–168, 1944.

[37] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[38] G. Y. Veroli, C. Fornari, I. Goldlust, G. Mills, *et al.*, "An automated fitting procedure and software for dose-response curves with multiphasic features," *Sci Reports*, vol. 5, no. 1, p. 14701, 2015.

[39] I. E. Naqa, G. Suneja, P. E. Lindsay, A. J. Hope, *et al.*, "Dose response explorer : an integrated open- source tool for exploring and modelling radiotherapy dose–volume

outcome relationships," *Physics in Medecine and Biology*, vol. 51, no. 1, p. 5719–5735, 2006.

[40] S. R. Gadagkara and G. B. Callb, "Computational tools for fitting the hill equation to dose–response curves," *Journal of Pharmacological and Toxicological Methods*, vol. 71, no. 1, pp. 68–76, 2015.

[41] D. J. Adams, "In vitro pharmacodynamic assay for cancer drug development : application to crisnatol, a new DNA intercalator," *Cancer research*, vol. 49, no. 23, pp. 6615–6620, 1989.

[42] R. Rahman and R. Pal, "Analyzing drug sensitivity prediction based on dose response curve characteristics," in *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on*, pp. 140–143, IEEE, 2016.

[43] J. D. Scheff, R. R. Almon, D. C. DuBois, W. J. Jusko, *et al.*, "Assessment of pharmacologic area under the curve when baselines are variable," *Pharmaceutical research*, vol. 28, no. 5, pp. 1081–1089, 2011.

[44] J. E. Rubnitz, B. Gibson, and F. O. Smith, "Acute myeloid leukemia," *Hematology/Oncology Clinics*, vol. 24, no. 1, pp. 35–63, 2010.

[45] F. Ferrara and C. A. Schiffer, "Acute myeloid leukaemia in adults," *Lancet*, vol. 381, no. 9865, pp. 484–495, 2013.

[46] D. Bonnet and J. Dick, "Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell," *Nat Med*, vol. 3, no. 7, pp. 730–737, 1997.

[47] A. Burnett, M. Wetzler, and B. Löwenberg, "Therapeutic advances in acute myeloid leukemia," *J Clin Oncol*, vol. 29, no. 5, pp. 487–494, 2011.

[48] L. Bullinger, M. Ehrich, K. Döhner, R. F. Schlenk, *et al.*, "Quantitative DNA methylation predicts survival in adult acute myeloid leukemia," *Blood*, vol. 115, no. 3, pp. 636–642, 2010.

[49] M. E. Figueroa, S. Lugthart, Y. Li, E. Claudia, *et al.*, "DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia," *Cancer Cell*, vol. 17, no. 1, pp. 13–27, 2010.

[50] F. R. Appelbaum, H. Gundacker, D. R. Head, M. L. Slovak, *et al.*, "Age and acute myeloid leukemia," *Blood*, vol. 107, no. 9, pp. 3481–3485, 2006.

[51] C. Leith, K. Kopecky, I. Chen, L. Eijdems, *et al.*, "Frequency and clinical significance of the expression of the multidrug resistance proteins MDR1/P-glycoprotein, MRP1, and LRP in acute myeloid leukemia : a southwest oncology group study.," *Blood*, vol. 94, no. 3, pp. 1086–99, 1999.

[52] C. Leith, K. Kopecky, J. Godwin, M. T, *et al.*, "Acute myeloid leukemia in the elderly : assessment of multidrug resistance (MDR1) and cytogenetics distinguishes biologic subgroups with remarkably distinct responses to standard chemotherapy. a southwest oncology group study.," *Blood*, vol. 89, no. 9, pp. 3323–9, 1997.

[53] M. Brune, S. Castaigne, J. Catalano, K. Gehlsen, *et al.*, "Improved leukemia-free survival after postconsolidation immunotherapy with histamine dihydrochloride and interleukin-2 in acute myeloid leukemia : results of a randomized phase 3 trial," *Blood*, vol. 108, no. 1, pp. 88–96, 2006.

[54] S. S. Shapiro and R. Francia, "An approximate analysis of variance test for normality," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 215–216, 1972.

[55] S. N. Gardner, "A mechanistic, predictive model of Dose-Response curves for cell cycle phase-specific and -nonspecific drugs," *Cancer Res*, vol. 60, no. 5, p. 1417, 2000.

[56] M. W. Skwarchuk, A. Jackson, M. J. Zelefsky, E. S. Venkatraman, *et al.*, "Late rectal toxicity after conformal radiotherapy of prostate cancer (i) : multivariate analysis and dose–response," *International Journal of Radiation Oncology, Biology and Physics*, vol. 47, no. 1, pp. 103–113, 2000.

[57] V. Guardabasso, P. Munson, and D. Rodbard, "A versatile method for simultaneous analysis of families of curves.," *Faseb J Official Publ Fed Am Soc Exp Biology*, vol. 2, no. 3, pp. 209–215, 1988.

[58] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, *et al.*, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, p. 603, 2012.

[59] R. Shoemaker, "The NCI60 human tumour cell line anticancer drug screen," *Nature Reviews Cancer*, vol. 6, no. 10, p. 813, 2006.

[60] D. B. Solit, L. A. Garraway, C. A. Pratilas, A. Sawai, *et al.*, "Braf mutation predicts sensitivity to mek inhibition," *Nature*, vol. 439, no. 7074, p. 358, 2006.

[61] J. Greshock, K. E. Bachman, Y. Y. Degenhardt, J. Jing, *et al.*, "Molecular target class is predictive of in vitro response profile," *Cancer Res.*, vol. 70, no. 9, p. 3677, 2010.

[62] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, *et al.*, "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, no. 7391, p. 570, 2012.

[63] M. Fallahi-Sichani, S. Honarnejad, L. M. Heiser, J. W. Gray, and P. K. Sorger, "Metrics other than potency reveal systematic variation in responses to cancer drugs," *Nature chemical biology*, vol. 9, no. 11, p. 708, 2013.

[64] M. Ramirez, S. Rajaram, R. J. Steininger, D. Osipchuk, *et al.*, "Diverse drug-resistance mechanisms can emerge from drug-tolerant cancer persister cells," *Nature communications*, vol. 7, p. 10690, 2016.

[65] D. A. Flusberg and P. K. Sorger, "Surviving apoptosis : life–death signaling in single cells," *Trends in Cell Biology*, vol. 25, no. 8, pp. 446 – 458, 2015.

[66] B. Efron, *Bootstrap methods : another look at the jackknife*, p. 569–593. Springer, 1992.

[67] H. Adèr, *Advising on Research Methods : A Consultant's Companion*. Johannes van Kessel Publishing, 2008.

[68] R. Kohavi, "A study of cross-valisation and bootstrap for accuracy estimation and model selection," *IJCAI*, vol. 14, no. 2, pp. 1137–1145, 1995.

[69] C. Robert, *Machine Learning, a Probabilistic Perspective*. The MIT Press, 2002.

[70] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[71] N. Nachar, "The mann-whitney u : A test for assessing whether two independent samples come from the same distribution," *Tutorials in Quantitative Methods for Psychology*, vol. 4, no. 1, pp. 13–20, 2008.

[72] N. Malo, J. A. Hanley, S. Cerquozzi, J. Pelletier, and R. Nadon, "Statistical practice in high-throughput screening data analysis," *Nature Biotechnology*, vol. 24, no. 2, pp. 167–175, 2006.